Evaluating Local Economic and Employment Development

HOW TO ASSESS WHAT WORKS AMONG PROGRAMMES AND POLICIES

OECD

LOCAL ECONOMIC AND EMPLOYMENT DEVELOPMENT

Evaluating Local Economic and Employment Development

How to Assess What Works among Programmes and Policies

OECD

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Pursuant to Article 1 of the Convention signed in Paris on 14th December 1960, and which came into force on 30th September 1961, the Organisation for Economic Co-operation and Development (OECD) shall promote policies designed:

- to achieve the highest sustainable economic growth and employment and a rising standard of living in member countries, while maintaining financial stability, and thus to contribute to the development of the world economy;
- to contribute to sound economic expansion in member as well as non-member countries in the process of economic development; and
- to contribute to the expansion of world trade on a multilateral, non-discriminatory basis in accordance with international obligations.

The original member countries of the OECD are Austria, Belgium, Canada, Denmark, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The following countries became members subsequently through accession at the dates indicated hereafter: Japan (28th April 1964), Finland (28th January 1969), Australia (7th June 1971), New Zealand (29th May 1973), Mexico (18th May 1994), the Czech Republic (21st December 1995), Hungary (7th May 1996), Poland (22nd November 1996), Korea (12th December 1996) and the Slovak Republic (14th December 2000). The Commission of the European Communities takes part in the work of the OECD (Article 13 of the OECD Convention).

© OECD 2004

Permission to reproduce a portion of this work for non-commercial purposes or classroom use should be obtained through the Centre français d'exploitation du droit de copie (CFC), 20, rue des Grands-Augustins, 75006 Paris, France, tel. (33-1) 44 07 47 70, fax (33-1) 46 34 67 19, for every country except the United States. In the United States permission should be obtained through the Copyright Clearance Center, Customer Service, (508)750-8400, 222 Rosewood Drive, Danvers, MA 01923 USA, or CCC Online: *www.copyright.com*. All other applications for permission to reproduce or translate all or part of this book should be made to OECD Publications, 2, rue André-Pascal, 75775 Paris Cedex 16, France.

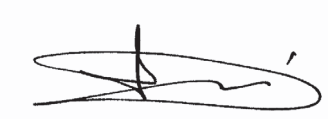
Foreword

A major challenge that faces public authorities responsible for local economic and employment development – and a critical challenge for policymakers wrestling with all forms of subnational development – is how to assess which programmes and which policies actually work. A corollary to this challenge is to identify, among the programmes that do work, those that provide the best value for money. In a macroeconomic context in which pressure on discretionary public spending is only likely to increase, not least because of the fiscal implications of the demographic transition, the need for answers to questions of policy effectiveness and efficiency will become all the more pressing. For a number of years now, and in a variety of fora, the OECD's Local Economic and Employment Development Programme (LEED) has drawn attention to the deficit in many OECD member countries as regards the volume and quality of evaluative research on the tools used to enhance local development. As part of its efforts to address the evaluation shortfall, the LEED Programme organised a major international conference in Vienna in November 2002 entitled "Evaluating Local Economic and Employment Development". This conference received generous financial and logistical support from the European Commission (DG Employment) and Austria's Ministry of Economic Affairs and Labour. The conference brought together many of the leading academics and practitioners in the OECD area concerned with such issues as: How do governments use the results of evaluative research? What is best-practice in evaluating the schemes that are often used to accelerate local economic and employment development? And can rigorous evaluation methods be used to measure the impact on entire localities of multi-instrument strategies and programmes?

Programme and policy evaluation raises issues that can be complex in conceptual and technical terms. However, an effort has been made to ensure that the papers are accessible to a non-technical audience. The papers focus on an array of programmes that have their principal impact on local labour markets and/or business development. Our hope is that these papers, and the assessment of policy implications set out in the opening chapter, will be of value both to the policy community and to those charged with the implementation of policies and programmes.

Improving evaluation practice, and building a more complete record of evaluation results, remains an ongoing priority of the LEED Programme. One idea that LEED will pursue is to compile an active on-line compendium of high-quality evaluation studies. Such a compendium could help to illustrate how certain perrenial evaluation challenges have been tackled in different circumstances by different institutions. In this

connection, the LEED Programme welcomes a continued exchange of views on all issues related to evaluation – and to local development practice more generally – with local authorities, academics and practitioners. Furthermore, in late 2003, the OECD established, in Trento, Italy, The Centre for Local Development. This Centre has a particular focus on the countries of Central and South-Eastern Europe and will have evaluation as one of the core components of its programme of work.



Sergio Arzeni
Head of the LEED Programme

Acknowledgement. *This publication is the result of a collaborative endeavour. Alistair Nolan and Ging Wong were responsible for all substantive aspects of the conception and development of the November 2002 Vienna Conference, on which the contributions to this book are based. Essential support in bringing the conference to fruition was provided by Jane Finlay, Jennah Huxley and Sheelagh Delf at the OECD Secretariat and Martina Berger at Austria's Ministry of Economic Affairs and Labour. Alistair Nolan has had overall responsibility for this publication, including editing most of the papers. Critical support in the production of this book has been provided by Sheelagh Delf.*

Table of Contents

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 1

Introduction and Summary

Evaluating Programmes for Local Economic and Employment Development: an Overview with Policy Recommendations

by
Alistair Nolan,
OECD,
and
Ging Wong,
Canadian Heritage and University of Alberta

The papers brought together in this volume were first presented at the conference "Evaluating Local Economic and Employment Development", held in Vienna in November 2002. This conference was organised by the OECD's Local Economic and Employment Development (LEED) Programme, with financial and logistical support from the European Commission (DG Employment) and Austria's Ministry of Economic Affairs and Labour. The holding of the conference was motivated by the widespread perception that there is a deficit in many OECD member countries with respect to the volume and quality of evaluative research on policies and programmes used to enhance local development. Why, for instance, is the evaluation literature on local development so relatively thin? Is this a result of inadequate public commitment to and practice of evaluation in this field, or perhaps a symptom of conceptual and methodological difficulties particular to local development? These and other issues were explored in the conference papers and discussions.

The conference attracted leading international figures in the field and sought to do three things: to consider how governments use evaluative research; to examine best-practice in evaluating the schemes most frequently used for local economic and employment development, and to consider whether rigorous evaluation methods can be used to assess the impacts on entire localities of multi-instrument strategies and programmes.

Use and misuse of evaluation

It is always the government's responsibility to ensure that public money is well spent, as alternative uses of funds constantly compete for policy spending priorities. The objective of evaluation is to improve decision-making at all levels – in management, policy and budget allocations. Evaluation is receiving renewed attention in OECD countries and is recognized as "important in a result-oriented environment because it provides feedback on the efficiency, effectiveness and performance of public policies and can be critical to policy improvement and innovation" (OECD, 1999).

Evaluation is essentially about determining programme effectiveness or incrementality, specifically the value-added of an operating programme or a potential public initiative. This primary purpose has become somewhat obscured by the fact that the work of evaluation has been largely focused on so-called formative evaluation activities, which provide information in improving

programme design and operations. Accurate information at this level is important but insufficient in a citizen-focused management regime that requires judgements of worth or merit. In this context, there is a growing demand for impact (sometimes termed "summative") evaluations (Canada, 2000). These are systematic attempts to measure the effects, both intended and unintended, of some government intervention, or mix of interventions, on desired outcomes. Such evaluative practices range widely in their complexity and rigour, often using comparative analysis across time, across participants and non-participants, or across detailed case studies. They typically rely on pre-post programme analysis ("single-system methods"), experimental or quasi-experimental designs, or detailed analyses of single cases that may be more feasible to apply in practice settings than in control-group settings. Design choices notwithstanding, such evaluations also require reliable and valid measures, as well as statistical procedures to determine the significance of an intervention on the outcome of interest. By establishing the links between stated policy, ensuing decisions, and impacts, evaluation provides an important learning and feedback mechanism, regardless of the specific moment of the policy process (State Services Commission, 1999). Building evaluation criteria into policy proposals forces an *ex ante* focus on desired outcomes. And *ex post* evaluation is an important tool for identifying policy successes and failures. Taken together, *ex ante* and *ex post* evaluations provide the critical evidence in support of results-based accountability.

Yet there is a low perceived demand for good evaluation of public policy in general and of local development in particular, depending upon the country and time in question. Numerous explanations for this have been offered, relating to both the production and uses of evaluation. On the production side, one is reminded of Henry Kissinger's reference to the heat of politics, in which *the urgent steals attention from the important*. Evaluation gets crowded out by other, immediate demands from ministers, especially against the background of fluctuating policy settings, the long timeframes needed for results to be realized, and the need to allocate funding to develop and sustain the necessary evaluation resources and technical staff capabilities (whether as evaluators or intelligent customers of evaluations). Different evaluation techniques also carry different price tags, with the gold standard of long-term random assignment experiments at one end of the spectrum, and process evaluations at the other. Where choices are forced, they are often in favour of the least expensive approach. Methodological issues also factor into policy managers' reticence towards evaluation. Evaluation against outcomes is considered just too hard:

> "… evaluation never provides uncontroversial answers. All social policy evaluation is plagued by the problem of the counter-factual – you never

have a control. All experience suggests it is expensive, difficult, and controversial." (State Services Commission, 1999).

Some also note the limitation of evaluation estimates of underlying population parameters or typical results. Viewed from the practical lens of public administration and public policy, one major issue with contemporary analysis is the overemphasis on average cases. While such analysis provides a great deal of useful baseline policy information, it is often not the behaviour of the typical and undistinguished that concerns us the most. More often, it is the exception to the statistical norm – the agencies that represent good or excellent practice, the identification of high risk programmes, or programmes that can meet multiple goals simultaneously such as efficiency and equity – that demands recognition or remedial attention (Meier and Gill, 2000). The focus on the high-performing or high-risk cases may highlight the reasons that separate them from the typical. The current state of evaluation practice does not handle such information demands well.

It is a long-standing observation that evaluation can be a double-edged sword in its uses. The very nature of a detailed scrutiny is a bottle half-full or half-empty, as even exemplary programmes have warts that may present politically-sensitive communication challenges to government. Concerns about how the results of evaluation might be used figure most prominently where self-interested stakeholders prevail – ministers and public servants might be equally attached to certain policies and reluctant to see them scrutinized. Leadership courage at the highest level is often needed to resolve such issues and to protect the evaluation function and its proponents. Thus in a number of jurisdictions there is the dilemma of nurturing an environment of transparency that avoids self-interest and capture while, at the same time, risk managing the evaluation function. Integrity is at the core and its observance or not will determine the level of public confidence in the evaluation function and its uses to improve accountability, management practices and budgeting.

Since the mid-1990s, sustained efforts to modernize comptrollership by a number of OECD countries have created some necessary conditions – and suggested mechanisms for government ministries – to develop a capacity for outcome evaluation. New performance reporting to government treasury departments increasingly demand *ex ante* information on how programme activities contribute to the achievement of outcomes, as well as *ex post* information on progress in working towards those outcomes. Budget processes today reflect incentives to encourage ministers to focus on continual priority-setting between low- and high-value policy results. At the same time, the field of evaluation is rapidly advancing in terms of the creation of rich panel data, new techniques and computing technology. It is on such a note of promise that we now turn our attention to the state of the art in evaluating local and regional development.

Evaluating local development

Throughout OECD Member countries significant resources are dedicated to programmes and policies to foster local and regional development. Bartik, in this volume, describes the magnitude of expenditures on subnational development in the United States. He cites an estimate that US$20-30 billion is assigned annually by local and state governments just to programmes of targeted business support. An additional amount of around US$6 billion is spent each year by the Federal government. These allocations take the form of direct spending on programmes and, overwhelmingly, tax incentives. Furthermore, such figures would be considerably enlarged – even doubled – if more general state and local government spending and tax incentives for business were included. In England and Wales, a 1999 study of local councils concluded that these spend "322 million pounds on economic development each year, and also manage billions of pounds of domestic and European regeneration funds" (Audit Commission, 1999). Public outlays on such a scale clearly merit a major investment in efforts to evaluate their effectiveness and efficiency.

Especially in poorer places, programmes and policies to promote local economic development encompass interventions in a wide range of sectors. Initiatives include actions in markets for property, labour, business information and advice, financial, health, education, and social services, policing, infrastructure, taxation, and institution building of different sorts. Many programmes have a long track record, in some cases stretching over a number of decades. The key features of such interventions are the subject of an abundant literature. The papers in this volume focus on an array of programmes that have their primary impact on local labour markets and/or business development.

By acknowledging the links and interactions between different interventions that have local development as their focal point, this volume addresses a departure from traditional programme evaluation. The conventional approach is to evaluate individual policy instruments and programmes against their explicitly stated objectives. In this way, programme evaluations tend to produce isolated and often disappointing findings, without due regard to the interaction and cumulative impact of policies that, by design or not, work in a "target-oriented" way (Schmid, O'Reilly and Schömann, 1996). This broader perspective recognizes that policies designed to influence area-based development do not exist in isolation and that an integrated approach is warranted.

There are too few high-quality assessments of local development policies and programmes

Given the magnitude of resources used for local economic and employment development, few countries have made a commensurate investment in generating rigorous evaluative evidence on which policies work and which do not. It is not simply that, frequently, there is limited funding of evaluation. In addition, as discussed below, there is considerable variation in the usefulness of many studies that purport to be evaluations. For instance, the local development literature is replete with case studies of local areas and their development programmes. However, many such studies simply describe a particular locality at a given point in time. They lack the longitudinal data on the area and/or its residents that might help to trace the causes of changes in economic or social circumstances. Consequently, policymakers are frequently unsure about which policy choices are best suited to which circumstances, and which policies should not be considered at all.[1] Indeed, the aforementioned study on councils in England and Wales found that many "are uncertain whether their efforts result in real opportunities for local people".

At least for locally-oriented programmes, evaluations often use methods and criteria that are insufficiently stringent to serve as a guide to policy. For example, there are around 1 000 business incubators in the United States. Most receive some public funding and many have local or regional development goals. There are numerous detailed studies of incubation schemes. However, to the knowledge of the authors, there has not yet been an assessment of business incubators in the United States that has used a control group methodology. In the absence of control-group assessments, ascribing changes in enterprise performance to the effects of incubation may be mistaken. Similarly, Boarnet (2001) shows that despite the popularity of "enterprise zone" policies, and the existence of a large number of evaluative studies, little of this research has been directly relevant to policy. This is because studies have been unable to identify which economic developments in the target areas would likely have occurred in the absence of the programmes. Indeed, across a range of programmes commonly used for local economic and employment development, too little is known about the **net outcomes** that can reasonably be ascribed to interventions.

Other difficulties also hinder the evaluation challenge. For instance, there is a tendency in many studies to examine (the costs of) output delivery rather than the achievement of net outcomes. That is, the focus of reporting is often on such variables as the numbers of training places provided, firms assisted, incubator units established, volume of credit disbursed, area of derelict land reclaimed, etc. The more complex question of how the supply of these outputs has affected the status of target beneficiaries usually receives less attention.

When relevant data are gathered, they are often collected at the end of a programme's life, with no baseline or *ex ante* information. Furthermore, it is not uncommon for evaluations to be produced by the very sponsors of the programmes being assessed.

Job creation is the most common yardstick of local development policy. However, studies often use opaque or non-standard measures of job creation. This makes cost-per-job-created claims unreliable and even misleading. For example, if a company has been created in, or moves to, a given location, and hires ten employees, this is invariably publicised as the creation of ten jobs. However, if the ten recruits simply moved from existing positions there may have been no net job creation (such redeployment might occur if jobs have been displaced from existing firms following competition from the new enterprise. Recruits might also have left old jobs voluntarily on account of more attractive conditions in the new openings). Typically, only a part of new hiring involves net job creation. But it is also possible that local welfare could rise even if all hiring involved redeployment. For instance, by comparison with the displaced positions the new posts might offer income or other gains to hirees. Conversely, recruitment could be associated with a decline in welfare if persons who are redeployed from displaced positions experience income or other losses. Reports of job creation do not always take such considerations into account.

Some studies also evaluate on criteria that are inappropriate. For instance, job creation measures can be unsuitable to assessments of the business development schemes that are a staple of local development strategies. The effects of such enterprise support schemes can be had on a range of business practices. They may impact, for instance, on the ability of entrepreneurs to adopt advanced management practices, to manage a company's inventory and cash flow, to raise product quality and lower process waste, to enter overseas markets, etc. Job creation can be a secondary effect of these outcomes, but need not arise automatically. More generally, many business development programmes enhance firm-level productivity. This can create pressure for labour shedding if demand for firms' output is static. Such considerations underscore the need to properly align evaluation parameters with the nature of the programmes being assessed. In the case of enterprise support, programmes often need to be evaluated on how specific business development practices in target firms have changed, rather than short-term impacts on job creation.

The overall paucity of high-quality evaluation is doubly regrettable in as much as local development initiatives are often intended to serve pilot or experimental functions. Policy piloting is, indeed, one of the claimed justifications for local development approaches *per se*. But this experimental function is squandered when programmes are not well evaluated. Indeed, in this connection, it is worth noting that some fundamental propositions about

local development and the policies that promote it are only poorly understood. For example:

- There is little quantitative evidence for the purported efficiencies of local development approaches *per se*. There are plausible generic reasons for thinking that local design and implementation should be superior for some types of policy and not others (for instance, superiority is unlikely in the case of policies that involve significant scale economies). But it is rare to find quantitative evidence of improvements in economic efficiency stemming from an increased role for the local level.
- There is limited understanding of the net effects of a range of business support schemes. For instance, as already mentioned, business incubators, despite their proliferation, have hardly been subject to systematic economic assessment anywhere. Similarly, in OECD Member countries, micro-credit programmes have rarely been evaluated using control-group techniques. And for this report only one systematic study was found that examined the local impact of self-employment support [see Cowling and Hayward (2000)].
- There is minimal information available that quantifies the costs, benefits and additionality associated with local partnerships, a frequent mode of programme design and (sometimes) delivery.[2]
- As Bartik notes in this volume, there is no direct empirical evidence for the notion that local employment benefits will be superior when there is a closer match between newly created jobs and the job skills of the local unemployed. Similarly, there is little empirical support for the contention that local employment benefits will be greater when local labor market institutions are more efficient in job training and matching.[3]

Furthermore – and not particular to the local development arena – there often exists a communication gap between policymakers and evaluation professionals. For instance, Eisinger (1995) showed that 34 out of 38 US states had conducted evaluations of at least parts of their economic development programmes in the early 1990s. However, only 8 states had made changes to programmes in the light of evaluation recommendations. Policy evaluation appeared to have little incidence on policy formulation.

All of the above does not imply that good studies are unavailable, or that development practitioners and policymakers are unaware of the issues at stake. The various chapters in this volume document valuable examples of a diverse range of high-quality evaluations. An early review, Foley (1992), points to a variety of careful studies that have wrestled with complex issues of programme deadweight, displacement, substitution and multiplier effects.[4] Particularly in the United States, numerous thorough assessments have been made of regionally-based science and technology extension services [see Shapira (2003) for a review]. Enterprise zones have been assessed with

increasing sophistication [see for example, Bondonio and Engberg (2000) and Boarnet (2001)]. Smith, in this volume, cites sophisticated evaluations that have addressed such varied issues as the effects of casinos on employment, and the growth impacts of public sponsorship of sports and sports facilities. Various national and subnational governments have also created guidelines and institutional capacities for evaluating local development programmes (see for instance, HM Treasury, 1995).[5] Among the counterpart organisations working with the OECD, including regional and local development agencies, ministries of labour and other public institutions, there is an intense interest in evidence about what works and why. Nevertheless, the OECD-wide picture is one of deficit with respect to the quantity and quality of policy-relevant evaluation. And the evidence base is weak even with regard to a number of the basic tenets of local development practice.

Why is there too little of the needed evaluative evidence?

There are a number of possible explanations for why the evidence base for policy is weak. These include the following:

- *Possible objections to evaluation among programme managers and implementing agencies*. This might stem from fear that support will be withdrawn if programmes receive a negative assessment. This is a problem for public policy evaluation *per se*. Objections might also reflect the fact that the more statistically sophisticated evaluations have often been useful in deciding whether a policy has worked, but have been weak in describing how the policy might be improved (Bartik and Bingham, 1997). Among programme administrators, this may have reduced the perceived usefulness of evaluation.
- *Practical and methodological challenges of rigorous evaluation*. Measuring such general equilibrium effects as deadweight, displacement and substitution is notoriously difficult. But evaluation of local development policies can involve additional complexity given that effects on a geographic area are being assessed in conjunction with effects on target groups (persons or firms). Effects on target groups need not translate into effects on the local area. This is the case, for instance, when created job vacancies are filled by in-migrants, or when an improvement in skills among residents facilitates their relocation, or when increased business activity leads to higher levels of out-of-area input procurement. Such difficulties are further compounded if evaluators have to consider how a number of policies interact across a geographic area that might contain multiple target groups. In addition, some local development outcomes can be difficult to quantify (such as reduced fear of crime, in the case of neighbourhood policing initiatives, or aspects of community capacity building).

- *The direct and indirect costs of evaluation*. Direct costs can be particularly significant for experimental or quasi-experimental forms of evaluation. Programme managers also sometimes view evaluation and monitoring as an intrusive source of administrative burden. In addition, evaluation can itself involve economies of scale (with certain fixed costs, for instance, in the collection and organisation of data) and of scope (as insights from one evaluation might be applied to others). Such economies imply that evaluations are often best sponsored and/or undertaken by higher levels of government. It is therefore unlikely that local authorities will produce a record of systematic evaluation that matches the extent of local policy innovation.
- *Incentives for local authorities to under-invest in evaluative knowledge*. Because the benefits from evaluation findings can accrue to agencies other than those that sponsor the studies, local bodies may under-invest in evaluation.
- *Policy and programme overload*. Evaluation can appear to be an unrewarding investment when, as is often the case, government initiatives are numerous and the population of active programmes changes constantly (a problem made worse when programmes have multiple objectives).[6] The extended time horizon over which some programmes yield measurable effects might also discourage policymakers from investing in evaluation.[7]
- *A weak understanding of evaluation techniques and principles, and a lack of suitably trained evaluators, in many local authorities* (a lack of in-house evaluation capacities can also place local authorities in a disadvantageous position when subcontracting evaluation studies).
- *In many countries, a lack of appropriate small-area data*. In some cases, the geographic units over which data is collected – in say health or education – do not coincide with the units across which the local development programmes act. In some contexts small-area data is simply unavailable.

Despite these disincentives, some OECD countries recognize the critical importance of systematic, rigorous evaluations in public decision-making. Canada, for instance, has made evaluation evidence obligatory for the renewal or reauthorization of all federal programmes. This comptrollership requirement is a powerful inducement to invest in and build an evaluation culture across the federal government. Such a national commitment has not typically been matched by subnational governments, which operate the majority of local development programmes. Yet the possible benefits from a greater quantity and quality of local development evaluations could be considerable. Most obviously, improved evaluation could help local and central authorities to allocate sizeable resources in an economically efficient manner. "Best-practice" in different programme types could be gauged in a meaningful way, such that different implementation modalities in given

programme types could be properly chosen. In addition, it is often observed that evaluation can be a learning exercise for evaluators and policymakers. Given that local development is a multi-sectoral endeavour, and that programme goals are sometimes vague or even inappropriate, a potentially important benefit of enhanced evaluation could result from encouraging implementing agencies to clearly specify what the goals of a programme are.

Careful programme monitoring is also critical and complex

Monitoring is a tool that can furnish information essential to programme management and superior programme outcomes. Monitoring can also ensure close contact with beneficiaries. However, monitoring is generally not equivalent to evaluation, being rarely concerned with issues of net programme outcomes. Taking the example of labour market programmes, Smith, in this volume, illustrates how performance standards need not (and generally do not) serve as a good substitute for impact estimates unless there is a systematic relationship between the two.

In addition to yielding information on programme implementation (and, possibly, impacts), performance indicators can define subcontract relationships with service providers and serve as an instrument of accountability. Furthermore, the choice of performance measures for programmes creates incentives that shape the ways in which services are provided. When the continued funding of programmes depends on the achievement of pre-specified performance targets, inappropriate performance measures can have serious and sometimes difficult-to-foresee effects on programme implementation and effectiveness. Accordingly, care is needed in the selection of performance indicators.

For example, an incentive exists to increase client throughput if the performance of a micro-enterprise scheme is assessed against the number of persons who enter the programme. Clients may be encouraged into the scheme – or be accepted when they should be dissuaded – regardless of probable business outcomes. In a similar fashion, when loan repayment rates have been used as the principal performance measure in micro-credit projects, staff have sometimes used legitimate accounting procedures to turn in high published rates of repayment, while programmes have been designed in ways that would preclude the reporting of low repayment rates (Woolcock, 1999). Similarly, if programmes are assessed against the number of enterprise start-ups they bring about, then support might shift away from services that are likely to enhance business survival. And funding which is based on output measures – while having the virtue of administrative simplicity – involves making payments after providers have incurred expenditures. This can cause support to be directed towards types of services that require little initial spending (Metcalf *et al.*, 2001).

In general, single-variable performance measures invite distortions in the running of programmes that have complex effects. Performance measures should be sought that reflect the complexity of the programmes and outcomes being monitored. For instance, using again the example of enterprise support programmes, performance measures could combine data on start-up rates – if enterprise creation is an underlying goal – with indicators of business survival – because simply creating firms is not enough. Higher weightings could be given to projects involving enterprises in which the size of capital invested and expected business incomes are comparatively high (it is in such firms that displacement is likely to be low). Similarly, to avoid the common bias towards working with entrepreneurs who would be successful even without the programme, higher weightings could be afforded to the establishment and successful management of firms by individuals who face barriers to enterprise (such as persons whose loan applications have been rejected by a bank).

Metcalf *et al.* (2001) note that service providers may face different performance measurement requirements from a range of programme sponsors. In this regard, governments can act to ensure consistency of performance measures across similar programmes. Bringing about a degree of standardisation in requirements for performance data can reduce administrative burden, especially for service providers that receive funds from more than one source. Such standardisation could also improve comparability across government-funded programmes, which would facilitate the generalisation of best-practice.

It is also important that monitoring not be perceived principally as a means of control. Service providers should be convinced of the utility of measuring performance. In practice, monitoring is sometimes performed in a perfunctory way. Service providers are often unconvinced that the data they are asked to assemble are useful.

Overview of the papers in this volume

Evaluation research and policy learning

Ging Wong's paper – *Policy Learning Through Evaluation: Challenges and Opportunities* – brings together a rich variety of insights from the former Director of the Canadian government's largest evaluation service. The paper situates the evaluation function within the broader context and processes of policy formulation, as well as providing a careful exposition of what evaluation is and is not. Clear distinctions are drawn between evaluation and financial audits, and between evaluation and various performance-based management and accountability systems. Wong concludes that the three fundamental tasks of evaluation are: to facilitate public accountability;

promote democratic processes, and enhance research and policy development. It is shown that the emergence of evaluation in North America, and its subsequent take-up in Europe, were directly related to the need for accountability in government expenditure budgeting. Wong also outlines the national and international institutional forces that have influenced the further expansion of evaluation in Europe. In this connection, he observes that the growth in demand for evaluation appears greater at European and national levels than at regional and local levels. Five ways are described in which evaluation can contribute to policy development. These involve improving:

- Programme design, through assessing the achievements of past programme performance.
- Programme implementation, through process evaluations.
- Programme cost-effectiveness.
- Programme management, through, for example, validating indicators and performance targets.
- Analytical and measurement capacities.

Evaluation is shown to augment knowledge of: needs and problems; effective practices and programmes, and programming.

Wong observes that the uses and relevance of different forms of evaluation – prospective, formative and summative – depend on the phase of policy development. Prospective evaluations, based on compilations of data or analyses, are particularly apt for early stages of policy formulation. Once a course of action has been decided, and a programme established, formative or process evaluations can help to identify and rectify problems of implementation. Lastly, summative evaluations seek to isolate the final outcomes attributable to the programme.

However, the paper points out that even well prepared and insightful evaluations might not play a major part in shaping policy. In practice, evaluative evidence constitutes only one input to the process of policy development. Political considerations and public opinion, for instance, can also play a role. Wong holds that policy development cycles are themselves becoming shorter, and that this is a source of pressure against the undertaking of rigorous evaluations, which are more time-consuming and expensive to prepare. Consequently, there is a greater reliance on less precise approaches to evaluation. A further problem area highlighted in the paper is that of how evaluators communicate their findings to a policy audience. Communicating technical material clearly to non-specialists and non-statisticians requires particular care.

Wong notes that evaluation is one of the few sources of reliable evidence on the achievements of policy, and that long-term government commitment to evaluation is essential. The challenge that Wong sees at the local level is one of building participatory evaluation strategies that involve the numerous stakeholders involved in local development. There is a parallel and related need for high-quality local case studies, performed in a multidisciplinary way, that can complement other forms of evaluation. Such case studies should serve as inputs to meta-studies done by higher levels of government.

Robert Walker provides an academic and critical practitioner view of *Evaluation: Evidence for Public Policy*, drawing extensively from recent United Kingdom policy evaluation case studies to illustrate three key challenges. The first considers the basic evaluative questions posed by public policy and links these to evaluation techniques for answering them. The choice of technique or evaluation model is seen to turn on the fundamental questions of whether and how a policy works and, equally, the time perspective (past, present and future) of the evaluation question. As Walker shows in his selection matrix, the techniques available are numerous and varied. They reflect, for the policy sponsor, different ways of doing the more familiar "formative" and "summative" evaluations associated with process implementation and outcomes measurement phases of the policy cycle. Walker is not unaware of these retrospective evaluation approaches to existing policies; rather, he acknowledges an increasing appetite for prospective evaluations to develop potentially new policies. Here, the current Labour government in the United Kingdom, following practice in the United States, is investing in building capacity to evaluate pilots, prototypes or demonstration projects before making final decisions on the design of new policies. Random assignment experiments and micro-simulation are favoured to test well-defined counterfactuals, while meta-analysis is used to systematically aggregate and summarize results from existing studies to identify successful aspects of policy and implementation.

The choice of evaluation instruments, however, is greatly influenced by changes in the policy environments. Herein lies Walker's main contribution – his critical insights on the institutional and political environments that have shaped public evaluation efforts. Recent British history presented both opportunities and threats to evaluation. Walker offers an explanation for the radical shift towards the greater use of evaluation evidence in policymaking over the last twenty years in the United Kingdom. This development was stimulated in the 1980s by the emergence of the "new public management", with its emphasis upon monitoring, control and performance measurement, and that became embodied in HM Treasury expenditure oversight directives. This increased demand for evaluation evidence was accompanied by a substantial practice of evaluation, albeit along retrospective and quasi-

experimental lines, rather than the prospective and random assignment orientation that was becoming the evaluation mainstay of the United States. The reasons explaining these differences in approach were, for Walker, "mainly structural". Unlike the United States, British policies are highly centralized in their uniform implementation, offering little variations with which to empirically test the policy counterfactuals. Further, British policy is less dominated by positivist economics than in the United States, and the social sciences are more influenced by social action research traditions rather than the quantitative, comparative group analysis that is the dominant framework in the United States and Canada. The pace of evaluation work was quickened with the election of the Labour government in 1997, with its commitment to modernizing policymaking, to evidence-based policy, and to "building systemic evaluation of early outcomes into the policy process". Between 1997 and 2002, some 70 policy pilots were initiated by central government departments.

At the same time, however, Walker argues that even with a strong political commitment to policy evaluation, the present British policy environment is not naturally supportive of evidence-based policymaking. Certain features of the policy marketplace frustrate the most effective design and use of policy evaluations, including: the politicization of evaluation findings for policy advocacy; constraining evaluation timetables to accommodate short-term policy imperatives; evaluating policy problems that extend beyond the remit of single departments; a hyperactive piloting of policies in all localities, which limits the number of independent control variables available; and the lack of cumulative learning (or policy amnesia) that accompanies incoming governments with no access to policy files created by predecessors.

In short, Britain has largely succeeded in integrating evidence and policy evaluation into the policy process, but whether current practice is sustainable is open to serious doubt. The prototype evaluation dominates at the expense of other strategies, and to date the results on impact and cost-effectiveness have been disappointing. As policymaking itself has not accommodated the requirements for good evaluation practice, there may be little motivation in academic communities to build capacity for such work. To have convergence, an evaluation culture needs to be developed and, in Walker's view, can be promoted in five ways: the full range of evaluative models should be employed to address issues pertinent to each stage of the policy cycle; evaluation evidence should be used in a non-political, objective manner; evaluators and policymakers should have more realistic expectations of research and policy requirements; policymakers should set lower expectations of innovative policy impacts; and it should be understood that evaluation is exceedingly difficult to do well and requires sustained investments and pooling in

theoretical and practical knowledge, as well as good data, methodological expertise and creativity.

Taking stock of the evaluation of local development

Timothy Bartik's paper – *Evaluating Impacts on Local Economies: What Has Been Done and What is Doable?* – examines two broad themes: the extent to which evaluation methods that use control groups could be employed to assess programme impacts on entire localities, and the evaluation of programmes to assist businesses in local communities.

Bartik draws attention to the need for policymakers to look beyond evaluation of the proximate goals of policy and programmes. The public benefits of various types of support programme need to be assessed. These include fiscal benefits and increased employment and/or earnings for the unemployed or underemployed. At the subnational level, the assessment of such outcomes can be approached by using regional econometric and simulation models. Such models are considered later in this volume, in the papers by Treyz and Treyz and by Wilson. Bartik notes that fiscal and employment benefits vary widely, depending on the particular demographic, economic, labour market, fiscal and social policy conditions found in each programme area. For instance, fiscal effects will in part reflect the extent to which programmes affect population in-migration relative to business growth. This is because, in the United States, businesses are generally net fiscal contributors, whereas the average household uses more public services than it pays for in tax contributions.

The exclusion of individuals or entire areas from programme treatment is inherent in random selection experiments. If, for evaluation purposes, area-development programmes were to exclude entire localities then this might be politically contentious. However, Bartik observes that the designation of programme resources to particular localities is often driven by political rather than objective economic considerations. The arbitrariness involved in the allocation of resources obviates ethical objections to excluding some places from support for the purpose of randomized experimentation. Bartik also considers the biases present when comparing area-wide development programmes. For instance, studies of tax and other business development and location incentives may systematically underestimate effects on local economic growth. This is because areas where such incentives are an important feature of policy tend to be those that are more likely to grow slowly even without the incentive programmes.

Bartik's paper also addresses the evaluation of programmes to assist business in poor communities. A distinction is made between schemes that assist all firms in a given location – as with some enterprise zone programmes –

and those that service only a subgroup of eligible firms. Finding a control group for the former type of scheme is almost automatically precluded for a local authority. Bartik's paper also provides a review of five generic techniques available for identifying the impact of a programme in the absence of a randomized experiment. A brief review is likewise presented of the assessments of the impacts of state and local taxes on business location and growth.

The paper notes that while survey methods can be valuable, they are more likely to be reliable when assessing the impact of support services, rather than financial assistance. In the latter case firms may be motivated to respond in ways that ensure the continuation of monetary support. Indeed, Bartik notes that as a condition for receipt of financial assistance, some programmes have even required that beneficiary firms state that the assistance was critical to a location or expansion decision. In such a context, responses to *ex post* surveys are unlikely to be reliable.

Daniele Bondonio presents a commentary on Bartik's paper and considers the techniques and methods that Bartik discusses in the context of the programmes funded by the European Union (EU). With particular reference to EU programmes to support business in EU Objective 2 areas, he argues that there is a need: *i)* to be clearer on what rigorous evaluation actually is; *ii)* to improve data collection; *iii)* to better incorporate evaluation needs into policy design; and *iv)* to exploit the heterogeneity involved in regionally distinct forms of programme design and implementation. Bondonio observes that evaluations of programmes co-sponsored by the EU structural funds usually only attempt to measure changes in the target areas or businesses. They rarely if ever seek to estimate differences between the observed changes and what would have occurred in the absence of the programme. And while regional economic models such as REMI or INPLAN can be used to estimate area-wide fiscal and employment benefits, they cannot provide valid results if they use unreliable measures of the impacts that programmes have on proximate dimensions of business activity. In other words, without rigorous impact evaluations of business support programmes the assessments of broader multiplier effects will be inaccurate.

Bondonio makes a number of important observations and suggestions on data issues. He notes that the need for improved systems of programme monitoring is a common refrain in studies of the EU structural funds. However, an additional need – for the purposes of rigorous evaluation – is for good quality data on non-assisted firms and areas. The evaluation of spatially targeted business incentive programmes could be improved if plant-level data were collected across small geographic units, especially if these data could be combined with information from employer records and socio-economic data on residents. Bondonio points out that in the EU, NUTS_3 areas are the smallest geographical units at which official and reliable statistics are

currently easily available. However, NUTS_3 areas are larger than many assisted areas – such as Objective 2 areas – which is a hindrance to area-based evaluation. Evaluation of the relevant EU programmes could be greatly facilitated by the creation of integrated statistical systems providing easily accessible data sorted by small geographic units that remain stable over time. Integrated EU data systems should also include registries of firms that receive assistance from any public source. This could help to avoid comparisons of EU-assisted firms with enterprises that simultaneously receive support from other public sources.

More also needs to be done to incorporate evaluation into policy design in a strategic way. Bondonio suggests that some programme assistance might be reallocated to areas that exactly coincide with geographic boundaries for which detailed statistical information is available. It is also emphasised that the variation in policy implementation and design across different regions in which Objective 2 programmes are implemented presents an opportunity for testing the effectiveness of different policy designs.

Using forecasting models

The paper by Frederick and George Treyz – *The Evaluation of Programmes Aimed at Local and Regional Development: Methodology and Experience Using REMI Policy Insight* – describes the REMI Policy Insight model, a regional economic forecasting and policy analysis tool used widely in the United States and other countries. This paper and discussions of *ex post* evaluation are linked through the fact that analysts may obtain certain data inputs for the REMI model from the outputs of programme evaluations, while micro-level evaluation can fail to capture important wider programme effects that could be quantified using macro-modeling. More broadly, the *ex ante* modeling illustrated here is part of a search for quantitative evidence in decision making, of which *ex post* evaluation is a continuation.

The REMI model is most frequently used to quantify a wide range of regional/local economic impacts stemming from economic development programmes (such as business attraction initiatives), transportation infrastructure investments, and environmental and energy regulations. A particular benefit claimed for the model's use is the identification of unforeseen programme or policy impacts. The model integrates input-output, computable general equilibrium and econometric techniques. Input-output structures track inter-industry relationships. General equilibrium parameters capture important long-term responses to price, cost and wage signals. And econometric techniques validate empirical bases in the model. The authors provide a detailed exposition of the model's structure and data input requirements. In using simulation models, and based on a long track record with the REMI model, the authors advise decision makers to ensure that the

following six features are present: 1) specification to local conditions; 2) a proper theoretical and structural foundation; 3) integrated general equilibrium, input-output, econometric and economic geography methods; 4) a set of input and output variables; 5) year-by-year results; and 6) a record of use for a large range of projects across different regions.

Robert Wilson provides a synopsis of the Treyz and Treyz paper and the associated conference discussion. He recalls that other macro econometric models have also been developed which can be used to help evaluate labour market interventions, such as the Local Economy Forecasting Model (LEFM) in the United Kingdom. An important discussion theme related to the complexities and minimum level of expertise required to use such models. Indeed, results can be sensitive to how such models are operated and the input assumptions used. Wilson notes that the models can be used to provide a useful counterfactual of what might have happened in the absence of a policy intervention. The key conclusion from the discussion was that such models, as one element in an overall evaluation approach, can yield important evaluation insights.

Evaluating area-based development in the United Kingdom

Brian Robson's paper – *Area-based Policy Evaluation in the United Kingdom* – provides a comprehensive overview of the development and evaluation of urban regeneration initiatives in the UK over the last thirty years. During this period the UK has witnessed shifts in emphases in regeneration, from economic, social, environmental and property-based approaches to the current focus on partnerships and co-ordination across fields of policy (such as health, education and housing). Robson describes the major investment in evaluation of area-based programmes since 1997. Just in the financial year 2002-03, more than £8 million were allocated by the Office of the Deputy-Prime Minister to evaluations of regeneration initiatives. Local regeneration partnerships have been required to evaluate and monitor their activities. Much evaluative research, and new neighbourhood-level data, has been posted on government websites. And sophisticated inter-disciplinary practices involving researchers and practitioners, as well as participatory evaluation approaches, have been adopted. In terms of the magnitude of dedicated resources, and the sophistication of the methodologies used, it is clear that the UK experience of evaluating area-based schemes holds lessons for other OECD member countries.

The paper affords a useful exposition of various of the conceptual and methodological challenges to evaluating area-based schemes. The difficulty of these challenges is made clear in the fact that most studies still focus on outputs rather than (net) outcomes. Furthermore, the problems of identifying interactions between strands of policy remain largely unresolved.

Nevertheless, ideas for practical steps to take can be gleaned: for instance, Robson cites one case study that examined areas that had bid for an initiative (City Challenge) but failed. These areas were compared with others whose bids had been successful, with inferences then being drawn about the additionality of the scheme.

There is a valuable discussion of how data issues constrain area-based evaluation. A new policy emphasis is seen to have been given to collecting data pertinent to local and sub-district geographic scales, and to collating information from different data sources. New approaches to data collection have also been adopted that can facilitate assessment of area-specific policy. For instance, Robson describes how the Department for Education and Skills has moved to gather address-based pupil data so as to assess changes in educational achievement and relate these to area effects. As in other countries, different government departments frequently employ different geographical units for data collection. Greater uniformity here could assist the evaluation task.

Important observations are made on the scarcity of evaluations that use longitudinal analyses to track changes in the labour market status (or other variables) of households and individuals. This is important given that policies often have impacts on an area that differ from the impacts on the original residents. An obvious example is when the creation of new jobs provides vacancies that are filled by in-migrants. Attempts at such longitudinal survey work are briefly described, while it is noted that tracking residents who have left a target locality can be problematic. The need for longitudinal assessment is underscored by the realisation that some regeneration impacts occur over long time horizons.

Robson comments that some hard-to-measure factors are often critical to programme success. Examples include managerial aptitude, leadership, and sensitivity to community issues. Identifying such factors can be done through surveys and qualitative methods such as discussion groups. Robson conjectures that a future direction of evaluation in the UK may involve greater attention to measuring less tangible variables, including the somewhat nebulous concept of social capital.

A persistent impression emerging from Robson's paper is that the difficulties of isolating area-wide effects are exacerbated by the volume of active initiatives. It seems that a critical area of emphasis should be on designing policy in a strategic fashion in order to circumvent some of the complexities of evaluation. It is unclear from the paper whether serious thought has been given to action along these lines in UK policy circles.

Jonathan Potter summarises, and in some points expands on, Robson's paper and the related conference discussion. Among other observations, he

notes that when support for local development takes the form of fiscal incentives it may be less conspicuous, and therefore less prone to evaluation, than in the case of schemes involving direct budget outlays.

Evaluating programmes of business assistance

The paper by Oldsman and Hallberg – *Evaluating Business Assistance Programs* – provides a lucid overview of issues in assessing programmes to support enterprise creation and development. Support programmes can affect business performance in varied ways. Thus the authors explain that a clear review and articulation of programme design is the essential first step in evaluation. Defining an explicit programme logic will help to establish the scope of the evaluation, the outcomes to be assessed, and the likely chains of causality. Important observations are made on programme outcomes and how to measure them. Clearly, the outcomes to be measured must be appropriate to the goals of the programme. Readers may consider this a self-evident observation. But in practice business development schemes are often charged with achieving outcomes – such as social development goals – that can be wholly inappropriate. For instance, especially in local and regional tiers of government, policymakers have turned to business incubation as a means of achieving a wide range of economic and social objectives, with job creation the most frequent goal of the publicly supported incubation schemes. However, contrary to common practice, the focus of incubation should be on enterprise development rather than employment growth (OECD, 2003) (employment growth will generally follow successful commercial outcomes, while job creation arising from incubation often occurs after tenant firms have graduated from incubators). Therefore, appropriate measures of incubator performance would record different dimensions of enterprise development. These dimensions might include, for example, the time that enterprises need to establish market niches or develop new products, or the adoption of advanced management practices. A focus on counting jobs might result in schemes being classified as unsuccessful when they could yield employment over time.

Oldsman and Hallberg note that once there is clarity on the outcomes to be achieved there can be more than one measure of outcome achievement. Often, there are trade-offs in choosing different measures of a given outcome. The characteristics required of outcome measures are seen to include relevance, validity, reliability, and practicability.

The paper contains a very accessible overview of the different evaluation techniques available, with useful commentaries on how these can be applied to business support programmes (it is noted that while the evaluation techniques are themselves well documented, it is only relatively recently that they have been applied to programmes of business assistance). The

uncommon observation is made that before-after studies of programme beneficiaries, while generally of little use for evaluation purposes, can be of value in certain situations. Such a situation might occur when the identification of a control group is precluded by the comprehensiveness of a scheme's coverage and when there is an evident and easily identified relationship between the programme and the outcomes in question (a case study is described that illustrates this observation). The trade-offs involved in the choice of the evaluation technique – in terms of cost, complexity, and strength of causal inference – are also clearly elaborated.

The recommendations made by Oldsman and Hallberg echo a number of those presented elsewhere in this book. For instance, evaluation requirements need to be taken into account in programme design. Baseline data and programme records should be collected. Multiple evaluation methods should be employed. And evaluations should attempt to demonstrate causality, discounting alternative explanations and explaining causal mechanisms. One of the few issues not given attention in this paper concerns the timing of evaluation. Deciding when evaluations are to be performed is important because different enterprise support programmes have different gestation periods. Some offer the possibility of almost immediate benefits. For instance, a business incubator could provide instant access to real estate for eligible firms. Other programmes might require months or years before change is evident. In particular, the job-creation impact of entrepreneurship support programmes tends not to be felt over the short-term. The time period over which an evaluation is performed might need to exceed the duration of participation in the scheme, depending on the type of programme being examined.

Concerning the impacts of entrepreneurship policies on local development more broadly, a number of additional points might be made. The first is that, by themselves, entrepreneurship development programmes are often too small relative to the local economy to have registered significant effects (Bartik and Bingham, 1997). In addition, as mentioned in previous papers, evaluations that successfully attribute outcomes to an entrepreneurship support programme are only a first step in assessing impact on the local economy. Evaluators need to go beyond the immediate effects of the programme and trace its income and employment multipliers as well as its fiscal consequences. Also, the fact that specifically local impacts can be mediated by economic developments separate from the entrepreneurship programmes – such as labour in-migration – might need to be considered in the evaluation logic. For instance, to determine the impact of an entrepreneurship scheme on local employment requires various steps (see Storey, 1990). For example, it would have to be established:

1. Whether at the time of the evaluation the full employment impact has been achieved (this may be difficult to assess, even by the firms themselves).

2. How many of the jobs resulting from Step 1 are jobs that would not have arisen without the programme?
3. How many of the jobs from Step 2 do not involve displacement of jobs in other firms?
4. How many of the jobs identified in Step 3 provide employment for residents or in-migrants? (in addition, policymakers will often wish to know, with respect to the employed residents identified in Step 4, how many have left unemployment and/or how many come from particular target groups such as the young or ethnic minorities. Policymakers may also be interested in whether these jobs are of short or long duration, and of high or low quality).

Evaluating labour market programmes

Labour economists in the United States have been heavily implicated in the evaluation of public employment and training programmes since the 1960s, when President Johnson renewed federal job training as an essential element of his administration's *War on Poverty*. Furthermore, the American emphasis on evaluating these labour market programmes is distinctly different from the usual European practice. Accordingly, this comparative OECD volume would be notably incomplete without an assessment of the role that evaluations have played in developing technical innovations and in shaping employment policy in the United States. For this, we turn to Randall Eberts and Christopher O'Leary. In their paper – *Evaluating Training Programs: Impacts at the Local Level* – the authors trace the evolution of the key features of American training policies over the past forty years, and document the associated evaluations, while comparing and contrasting federal and state evaluation practices.

The main findings can be summarized as follows. First, while most federally-funded programmes are regularly evaluated to determine programme renewal, reauthorization or the design of new training initiatives, state supported programmes have received far less scrutiny. The reasons for this gap are: *a)* the added complexity of evaluating state and local programmes which commonly have a multiplicity of objectives, a prime example being customized training for firms as a way to promote local and regional economic development; *b)* policy advocacy pressures that diminish the political commitment to evaluation; and *c)* resource constraints. By contrast, Eberts and O'Leary note that evaluations have become an integral part of many state and local welfare reform initiatives during the 1990s. This has been a result of the federal agreement to grant state waivers to change aspects of federal programming conditional upon evaluations of social programmes. Much of the current knowledge of the incentive effects of different welfare reforms comes directly from these local evaluations.

Second, the two most popular assessment techniques for job training are performance monitoring, which tracks gross outcomes, and net impact estimation, which assesses the value added of an intervention. Training evaluations in the United States are essentially seen through the lens of net impact evaluations which compare mean outcomes of a representative sample of programme participants to a similar sample of non-participants. The evaluation design is typically by random assignment of individuals or a comparison sample selection based on observable characteristics in a quasi-experiment. In the case of job training, such evaluations also consider and adjust for an over-estimate of net programme impacts due to the "Ashenfelter dip" observed in earnings prior to job separation for dislocated workers. Seven criteria for good estimation methods are described and listed in their order of importance. A further discussion is provided on the practical issues of evaluating job training, including sample sizes, site selection, sample selection, survey implementation and data preparation for analysis. Key evaluation results on what works for whom and by how much are then presented for the succession of employment policies beginning with the Manpower Development Training Act (MDTA) of 1962, the Comprehensive Employment and Training Act (CETA) of 1973, the Job Training Partnership Act (JTPA) of 1982, and the Workforce Investment Act (WIA) of 1998.

Third, a large number of states subsidize worker training by providing customized training programmes for local businesses, not principally to improve the employability of workers, as in federal programmes, but as a means to promote the location and retention of businesses. Consequently, state or local evaluations are complex undertakings that ideally combine the individual worker net benefits of training with the net effect of training on productivity and wages and, in addition, broad regional indicators such as job creation and poverty reduction. All this introduces additional complexities and imprecision into the evaluation. To minimize these difficulties, Eberts and O'Leary identify three basic types of evaluation methodologies that may be appropriate for evaluating state customized training programmes: constructing comparison groups based on firms not individuals; using econometric models to explicitly evaluate the key relationships between the intervention and the intended outcomes, and using a representative firm approach to estimate the relative effects of programmes on a firm's financial status. Examples are given to illustrate these alternative evaluation strategies. The overall conclusion is that the evidence on state-financed job training programmes is very limited and, therefore, our understanding of training effectiveness at the local level is spotty and imprecise. That said, there is some evidence in a handful of studies of specific state training programmes to suggest that subsidies to private training increase training and improve

productivity for both firms and workers, while increasing wages, improving the work environment, and expanding the workforce.

Jeffrey Smith, in his paper *Evaluating Local Economic Development Policies: Theories and Practice,* provides a complementary but quite different point of reference to methodological issues raised by Eberts-O'Leary and Bartik. He argues that the current scholarly literature on how to evaluate programmes has made significant advances in the past fifteen years but that evaluation practice remains mired in the 1970s. In providing a practical, relatively non-technical guide, his purpose is to make the scholarly literature more accessible.

The plan for Smith's discussion centres on five key themes in evaluating local economic development programmes: *a)* the choice of econometric evaluation estimators; *b)* how differential programme impacts for individuals, groups, firms and localities affect both evaluation practice and our ability to think about evaluation design and interpretation; *c)* concerns about the implications of general equilibrium effects for policy evaluation; *d)* when not to do an evaluation, particularly for smaller programmes; and, finally, *e)* evaluation as a method for ensuring that programme operators serve the taxpayers' interest and public good rather than simply funnelling money to politically influential stakeholders, with local economic development justification as cover.

Specifically, Smith draws attention to the following insights from the literature for advances in methods and in practice. First, there is no magic bullet. Each category of econometric estimator provides the correct answer only under certain combinations of available data, how programme participation takes place, and parameters of interest. Instead, to choose and interpret econometric evaluation estimators it is essential to have an evaluation plan that maps out data, institutions, and parameters of interest, while taking into account heterogeneous treatment effects. This would ensure that estimator selection adheres to strict rules. Second, as local development programmes generally aim to create general equilibrium or multiplier effects, the choice of estimator to use and the unit of analysis for evaluation will miss or be biased by general equilibrium effects.

Evaluating programmes in the third sector

Andrea Westall's paper – *Evaluation and Third-Sector Programmes* – surveys emerging approaches to the evaluation of programmes implemented by not-for profit service organisations and so-called social enterprises.[8] Such "third-sector" organisations have come to play an increasingly prominent role in initiatives for local development. Their best-known function is perhaps as an intermediary labour market institution, providing work experience and

training in order to facilitate transitions to work in mainstream labour markets. Evaluation of the third sector is important because of the growing public engagement with such organisations. It is also important given that such bodies have thus far been the object of little careful assessment, while they perform functions that have in some cases presented innovative responses to local economic and social problems. Furthermore, third-sector organisations themselves often wish to accurately gauge the full extent of the benefits they create and improve their service delivery.

Third-sector organisations are frequently accustomed to monitor and report on inputs, such as volunteer hours worked and the amounts and sources of financial contributions. Many collect information on outcomes among clients. But there is a relative unfamiliarity with more sophisticated impact assessment techniques, including assessments against pre- and post-programme data, client sampling procedures, follow-up with clients who leave a programme, demographic differentiation of data on beneficiaries, and the use of comparison groups.

Westall holds that under-evaluation of third-sector programmes in part reflects internal capacity constraints as well as the types of (output-based) monitoring and evaluation criteria used by programme sponsors. Diversity of reporting requirements – to multiple programme sponsors – is also said to have hindered evaluation. The superficial character of some evaluations might in part reflect the fact that such programmes are rarely funded by central authorities – the usual proponents of rigorous forms of assessment. Furthermore, because the funding volumes for third-sector schemes are small, the use of more stringent and costly evaluation techniques might not be worthwhile.

The existence of internal capacity constraints on evaluation is partly a financial problem. This highlights the need to earmark funds for evaluation in public budgets allocated to third-sector bodies. Such funds could help to cover outlays on staff time, mailings, telephone surveys, etc. National and local public and non-governmental organisations might also organise technical assistance to upgrade evaluation capabilities in third-sector bodies.

Considerable store is set on the potential advantages of involving community members and beneficiaries in the evaluation process. Westall holds participation in evaluation to be important, both as a means of gleaning information and as a way of creating a sense of programme ownership. Potential drawbacks are recognised, but perhaps not given sufficient weight. Westall notes that shortcomings include the resources that organising participation might require, and the danger of biased responses if the views of more vociferous or influential residents predominate. However, in more general terms, there is a danger that participatory approaches (if participation

means more than just being a survey respondent) might conflict with the objectivity that must underlie evaluation. Furthermore, in this volume Smith notes that little direct evidence is available on the reliability of participant self-evaluation. He refers to findings from experimental evaluation and behavioural decision theory that suggest that most people are not proficient in answering the sorts of counterfactual questions that evaluators need to pose. Woller's paper touches on a number of additional weaknesses in participatory approaches. These include a lack of standardization in the techniques used (although this need not be the case if central bodies work to ensure common procedures and formats), and problems stemming from disparate interpretations of impact within the same target group. As Woller notes, participatory techniques cannot claim to establish causality. But through methodical cross-checking of different forms of evidence a credible case for causality can at least be proposed. Westall's emphasis on combining methodologies – more rigorous techniques with (participatory) surveys – is reiterated in a number of the papers in this book.

Evaluating territorial employment pacts

The practical experiences of evaluating area employment pacts by government ministries were the subject of a workshop session. This workshop featured two papers – Paolo Casavola's *Evaluating Territorial Employment Pacts in Italy* and Peter Huber's *Evaluating Territorial Employment Pacts in Austria* – as well as a summary and commentary on the discussion in Hugh Mosley's *Evaluating Local Employment Pacts in Austria and Italy.*

Casavola introduces the concept of a territorial employment pact as "a specific policy instrument aimed at promoting local development through financial incentives to a group of locally based and integrated projects designed by a coalition of local actors (private and public)". While individual territorial pacts do not cover large areas, they are extensively used in Italy, with 230 in place in September 2002. Most of these were selected through a national procedure and a small number selected by an established European Commission procedure. In theory, such territorial pacts can accelerate the location of value added activities by promoting economies of agglomeration and other synergies. To evaluate this policy, Casavola distinguishes two separate but related issues – mechanisms for inducing local development (project selection and public-private coalition capacity) and final economic outcomes. In the Italian case, formal evaluation – both formative and summative – is not in place to address the necessary information and accountability requirements. In part, this reflects the administration's view that pacts should be in operation for sufficient time in order to assess results. In the interim, administrative monitoring of financial contributions to the pacts has been established. Academic studies and fieldwork have also been

commissioned to build territorial statistics associated with pact coverage and to conduct surveys of stakeholder entrepreneurs and key informants. Based upon these data sources, Casavola reports preliminary evaluation findings on the OLS regression of ongoing pact performance (effective expenditure rate of the public financial contribution) and explanatory variables for type of pact, initial economic conditions and social capital. Stakeholder participation appears to be a significant factor in quicker set-up and implementation.

Huber outlines the approach chosen to evaluate territorial employment pacts in Austria. Since the introduction in 1997 of four Austrian pacts by the European Commission, with the aim of combating unemployment, nine additional provincial-based pacts were created through a national action plan to support such territorial agreements by special subsidies. These provincial pacts varied in their purpose, from coordination of provincial labour market and economic budgetary measures to the coordination of policy innovations. Early studies and external evaluations of the structure and development of pacts are discussed, as well as efforts at self-evaluation. A particular problem for evaluation is the paucity of meaningful administrative data associated with coordinating activities. The pacts, however, produce regular management reports which can be a source of information to generate indicators on pact development. Given these circumstances, Huber recommends an evaluation strategy that focuses directly on processes – a formative evaluation that assesses policy formulation, implementation and uptake. This orientation also emphasizes whether pacts have contributed to establishing social capital in the region. This remains an empirical question for the formative evaluation to answer.

Mosley puts the evaluation problem into perspective when he reminds us that pact networks are not formal organisations with their own budgets and employees. Rather, they are voluntary partnerships in which the constituent members retain their own identities. Therefore, one value of such networks is the leveraging of a disproportionately large amount of joint activity that otherwise might not exist – that is the nub of the evaluation question or the counterfactual to be examined. The presumed sources of value-added are synergy effects and social capital formation associated with favourable long-term impacts on local employment and economic development. In answering the question "what's different about evaluating territorial pacts?", Mosley provides a useful point of departure for how to think about the evaluation requirements for all local development policies. In the first instance, and in contrast to individual- or firm-oriented policies, the central evaluation question is whether the local or area effect is achieved. A second evaluation question is the effect of governance arrangements on the level and mix of local policy. Another distinctive evaluation question is the impact of programme-induced changes in governance on aggregate policy outcomes in

the regions. A control/treatment group analysis is inappropriate here and an alternative aggregate impact analysis is suggested instead that might utilize a time series and/or cross-section framework, or a matched comparison of regional units. The final and classical evaluation question is the programme impact on individual participants. Given the heterogeneity of territorial pacts and the assumption that local implementation matters, a locally-focused evaluation is required, ideally based on local labour market outcomes for participants and a comparison group. A prime consideration here is whether economies of scale make evaluation of programmes in some localities too costly. Leitner's study of the Vienna Territorial Employment Pact demonstrated the technical feasibility of a rigorous evaluation using a quasi-experimental design based on a propensity score matching procedure for constructing participant and non-participant comparison groups. Mosley also noted other preconditions for success in undertaking such an evaluation – evaluation is a mandated activity for the European Social Fund, cost was not an obstacle in the case of Vienna, evaluation expertise was available, microdata was readily available from social security records, and sufficient time was allowed to conduct a proper evaluation.

Evaluating microfinance

Gary Woller's paper – *A Review of Impact Assessment Methodologies for Microenterprise Development Programs* – affords a comprehensive review of evaluations of micro-enterprise schemes both in OECD and less developed economies (LDCs). In total, 88 evaluation studies were considered. These consist of 67 separate impact assessments in 31 LDCs and 20 evaluations in 2 OECD countries, 19 of which were in the United States.

Woller notes that microfinance schemes have impacts that can affect the individual, the enterprise, the household and the community. A spectrum of economic, social, policy and location-specific factors impinge on these impacts. The paper provides a typology of the evaluation approaches that have been used to assess microfinance, from random experiments through a continuum of decreasing rigour to market research aimed at informing management decisions. The merits and demerits of each approach are reviewed in turn. It is noteworthy that of the 20 microfinance evaluations in OECD economies, only seven used a control group. This compares with 73 out of 90 of the evaluations of programmes operated in LDCs. Woller observes that while the absence of control groups precludes causal inference about programme outcomes, assessments without control-groups can still help to achieve certain managerial functions. They can, for instance, monitor client progress, assess relative outcomes among different market segments, and calculate unit costs for specific programme outcomes or outputs.

Another interesting issue raised in this paper relates to non-random programme placement. Programme managers may site a scheme on the basis of considerations that could have an independent influence on programme performance. For example, good logistical conditions, that affect whether managers can easily visit a scheme, could influence the ability of micro-firms to market their output. This could lead to programme outcomes different from what would have been achieved had the same programme been situated elsewhere. To counter this problem, programme placement might itself be randomised. Given that potential host locations for a programme are likely to exceed programme capacities, randomizing where the next scheme is to be established need not be impractical. Woller cites research from an evaluation in Northern Thailand which demonstrated that not accounting for non-random programme placement (and participant self-selection) led to significant overestimation of impact.

Loan fungibility is shown to be a particular problem facing the evaluation of micro-finance schemes. For instance, loaned funds might be used for purposes other than enterprise development, while loan repayments might not be made out of enterprise cash flows. Loan fungibility compounds the problem of drawing a direct connection between the receipt of a loan and changes in borrower/household income, consumption, asset accumulation, etc. The paper notes that to tackle this issue survey questions might attempt to acquire information on the uses of loan funds and the sources of loan repayments.

Woller makes important observations on how the presentation of evaluation findings can mislead. For instance, the loan repayment rate is a commonly used indicator of portfolio quality. However, this measure can omit important information contained in a superior indicator such as the share of the portfolio at risk. Assessments also often neglect to report on numbers of programme dropouts. Nevertheless, the rate of dropouts can be a robust indicator of whether participants feel the programme is worthwhile. Enterprise income can also be misreported in a variety of ways. For instance, the level of enterprise income is often described instead of changes in income. Similarly, income is sometimes reported without reference to how this income relates to time worked and capital invested. What the measured income actually comprises might also be unclear. In response to these problems, and as a general principle in commissioning evaluations, evaluators should be required to fully disclose the methodologies used. It should be a stipulation that methodological drawbacks in the chosen approaches, difficulties encountered, the rationale for the choice of performance indicators, areas of subjective assessment and possible conflicting interpretations be made explicit. As Woller emphasizes, full disclosure is critical for informed interpretation of evaluation findings by policymakers and others who might not be evaluation specialists.

Panel debate

Alice Nakamura's paper describes the discussion at the Conference's final debate, which brought together six eminent panelists (Professor Edward Hill, from Cleveland State University, the Right Honorable Henry B. McLeish, former First Minister of Scotland, Dr. Stephen Wandner, from the US Department of Labor, Professor Philip Davies, from the UK Cabinet Office, and Professor Alice Nakamura, from the University of Alberta). The exchanges focused on government commitment to evaluation and how to improve on the current situation. The issues treated ranged from the choice of indicators of economic development, to tensions between policy makers' tendency to guard the details of evaluation processes, the links between policymaking and academia, the open-access character of certain evaluation practices in the United States, possible incentives for the encouragement of good evaluation practice, the integration of data collection, methodological practice, policy design and implementation, and ways to improve the communication of evaluation information to the mass media.

Key themes emerging from the conference

Strategic design of evaluation

It is clear from the papers, and the wealth of practice that they describe, that few governments make use of the heterogeneity involved in regionally/locally distinct forms of programme design and implementation. Variation in policy implementation and design across regions and local areas – and randomised programme placement – present an opportunity for testing the effectiveness of different policy designs. Clearly, more needs to be done to incorporate evaluation into policy formulation in a strategic way. Even at the level of individual programmes there is ample scope for generating policy-rich information through local variation in implementation modalities. For instance, a regional authority sponsoring a micro-credit programme could vary key features of programme design – such as loan repayment schedules and group or individual borrowing techniques – across a number of local areas. In a similar strategic manner, but with a view to improving data availability, Bondonio's paper suggested that the EU might reallocate some programme assistance to local areas that exactly coincide with areas for which detailed statistical information is available.

The alleged intangibility of important local development outcomes

During the Vienna conference the view was often expressed that many local development outcomes are intangible or hard to quantify. Indeed, this contention is frequent among local development practitioners. For example, it

is sometimes held that third-sector agencies are able to create a range of beneficial outcomes that can be hard to calibrate. Such benefits are said to include enhanced social capital, greater community engagement with regeneration initiatives, self-confidence among participants and the generation of (superior) local information. But the purported intangibility of local development outcomes may have been overdrawn. In the first place, some outcomes might be hard to measure principally because their conceptual bases do not yet command a consensus. Such is the case with the contentious concept of "social capital". The evaluation challenge here might relate first to whether the programme sponsors had a meaningful concept of social capital: that is, whether the programme logic was coherent. Programmes that seek to enhance "community cohesion" might not be amenable to evaluation not because some components of cohesion cannot be measured, but because the authors of the programmes had not specified what was meant by community cohesion at the outset. Other types of less tangible benefit might in fact be measurable. For example, self-confidence and local opinions of development initiatives can be gauged through surveys. When an outcome can be stated specifically it can often be measured in some way, whether directly or by means of proxies.

Tensions between "scientific" and other approaches to evaluating local development

A number of the papers, and various interventions during the conference, suggest a false opposition in some quarters between econometric and participatory approaches to evaluation. The two should represent points on an evaluation continuum, with both affording insights of different kinds. Woller's paper describes the critiques of scientific techniques used to evaluate microfinance. These are seen to include assertions that such methods: fail to capture the full complexity of causal relationships; attempt to quantify what cannot be quantified; empower technocrats; and fail to result in action by or for poor groups. Not all of these arguments are compelling. For instance, it is hard to see why the use of scientific methods is more likely to involve misunderstanding of causality than other forms of evaluation. Statistically literate analysts are at least as likely to be aware of complex causal interactions as analysts with strengths in other forms of enquiry. However, the claims made against so-called scientific methods do signal to researchers a need to be alert to the wider context in which a programme is being assessed, and to give due consideration to evidence from programme participants.

Displacement and the need to evaluate what is of greatest policy relevance

Displacement was a theme much discussed during the conference. Displacement takes various guises and is present in a number of aspects of local development policy. For instance, there is a concern that countering criminal activity in one locality might lead to its displacement to another. And a partnership project might displace resources – such as the time of individual participants – from other uses.

There are severe methodological problems in gauging the magnitude of displacement. To take the case of enterprise displacement: a particular complication is that the measurement of displacement usually depends on surveys of the firms that may have caused displacement in other firms. The displaced enterprises are sometimes no longer in business, and the views of the two groups of firms could differ systematically. In addition, there is no single standard rate of displacement: a variety of context-specific conditions determine the severity of displacement. For instance, displacement is likely to be greatest among the types of firm that predominate in poor localities. These are firms in mature low-growth sectors in which skill and capital requirements are limited and in which barriers to entry are low. Furthermore, the state of local demand and supply is critical to the magnitude of displacement effects. And while displacement is likely to be most acute at the local level, it need not be limited to the confines of a particular locality: firms elsewhere can also be displaced. Also relevant to displacement is that there are considerable differences in the geography of markets served by firms in different sectors (service sector firms are more likely to have a local orientation, especially those that provide personal services). Displacement will also vary depending on whether the products of firms in the same sector are close substitutes or not. In the light of these observations, it is unsurprising that estimates of displacement have varied significantly [right up to one hundred per cent in some enterprise support schemes (OECD, 2003)].

In summary, it is known that the complexity of the phenomenon suggests that estimates of displacement be treated with caution. It is also known which stylised conditions are likely to be associated with higher or lower levels of displacement. Accordingly, further policy-related evaluation is perhaps most needed in the assessment of measures that might mitigate displacement. For example, programmes to help firms to sell in out-of-area markets, and to raise the average size of investment in new businesses, may both reduce displacement (OECD, 2003). Such research might yield at least as much policy-relevant information as ever-more-detailed attempts to pin down the scale and extent of displacement, which are in any case elusive and highly context-specific.

Recommendations for policy: rationale

At the Second OECD Ministerial Conference on small and medium-size enterprises (SMEs), in June 2004, policy evaluation was recognized as key to supporting innovative SMEs. A contributor to the Conference made the following observation, which is in fact applicable across the full range of public policy:

> "Policy makers should be able to answer four questions about their policies. Are they clear and coherent? Do they have clear objectives? What are the targets? Are they being evaluated? If you don't have evaluation, you're kidding yourself that you're achieving your objectives." (David Storey, CORDIS News, 2004-06-07)

In practical terms and based on the expert advice presented in this volume, the implementation of the following recommendations will contribute to an evaluation culture that will produce better quality and more useful evaluations to inform decisions on local development.

For central and local governments

Streamline and rationalise the development and implementation of programmes and policy

The evaluation task is greatly complicated when a shifting population of government programmes overlap. When programmes are replaced prior to evaluation, or are evaluated before their full impact can become evident, public resources may be squandered. And when initiatives are too numerous, investing in policy learning can appear unproductive.

Make explicit, at the highest possible level, the commitment to and importance of evaluation. There should likewise be a commitment to public diffusion of evaluation findings

An overt recognition of the importance of evaluation, by senior policymakers and agency heads, is vital. Human and financial resources for evaluation are more likely to become available once such recognition is evident. A proportion of overall programme budgets should be earmarked for evaluation purposes at programme inception. For example, in the UK, for a number of regeneration programmes, a share of the available funding has been set aside for evaluation (DETR, 2001). And all Australian public policy proposals require the inclusion of an evaluation strategy.

To encourage greater openness to evaluation, it should be made clear that the aim of evaluation is to improve the quality of public policy. Regular government reporting of performance should draw extensively from evaluation results. Evaluations that reveal problems in a given programme

should not be seen to provide automatic grounds for termination of the programme. Rather, evaluation should be viewed as a tool to provide a basis for improved policy. Furthermore, the dissemination of evaluation studies and results – possibly on an anonymous basis – can serve as a form of evaluation quality control and as a means of stimulating policy-relevant academic analysis.

Governments should also commit to placing raw evaluation data in the public domain in ways that are affordable and accessible for potential users. For example, as part of evaluations of a number of labour market programmes sponsored by the US Department of Labour, primary data have been made available on-line at low cost.

Governments could likewise create or support institutions that provide training and technical assistance in evaluation, especially to subnational authorities, and/or serve as centres of evaluation excellence. Governments could also establish or support fora (conferences, seminars, workshops, Internet sites, etc.) that promote debate and the dissemination of information on evaluation results and methods. An important function of evaluation fora or centres of excellence could be to foster understanding between the research and policy communities.

Mandate evaluations when public funding is provided

A key reason why various state and federally sponsored programmes in the United States have been thoroughly evaluated – such as the Manufacturing Extension Partnerships programme – is because mandatory evaluation requirements were attached to the use of federal government funds. Central (or regional) government requirements for evaluations could also involve standardised data collection across subnational agencies (or build standardised data collection protocols into programme design). This could help to compare the impacts across different local settings of some widely used programmes.

Clarify the outcomes to be evaluated during programme design, and certainly prior to programme implementation.

Prospective programme managers should be asked to detail the steps required to evaluate the programme's impact, including the collection of data for evaluation purposes. The preparation of a blueprint for the evaluation process would ideally also involve the evaluating agency. Issues that might be clarified include the choice of alternative methodologies and the trade-offs involved, the types of impacts to be examined, the types of data to be collected, the indicators to be used, and how the results would be reported. Deciding when evaluations are to be performed is also important, as different

programmes have different gestation periods. Preparing for and undertaking good-quality evaluation is itself a pedagogical exercise which has the potential to improve thinking about programme design. For instance, local development initiatives often involve multiple objectives (social, environmental, economic, etc.). But the simultaneous achievement of multiple objectives can be unrealistic. The prior specification of objectives could help to avoid investment in inappropriate programmes.

Evaluate when there are likely to be net benefits from doing so

Some types of programme have been evaluated in rigorous ways on multiple occasions in a number of countries. New evaluations of similar programmes may not represent a good public investment. Smith, in this volume, notes that "Time spent reading the literature for good evaluations of similar programmes may yield more useful results than a weak evaluation based on poor data completed by a poorly qualified evaluator using inappropriate methods." There is also obviously little sense to investing significant resources in evaluating programmes that will in any event be discontinued. Furthermore, for small programmes, the potential benefits obtained might not justify the cost of sophisticated evaluations. Similarly, evaluation may be superfluous if the available data is of poor quality and/or if sample sizes are too small to allow statistically valid results.

Choose the evaluation technique in the light of the size and nature of the programme concerned

Studies of major programmes should use a variety of methods: random assignment, quasi-experimental assessments, interviews with beneficiaries, participatory approaches involving all stakeholders. Assessments of outcomes using experimental evaluation approaches are expensive. They are also data intensive and require particular statistical expertise. As such, these techniques are often not appropriate to lower levels of government that operate relatively small programmes. When applied by subnational governments, they should be used to assess the largest programmes (or pilot programmes that might be significantly enlarged) where the potential benefits from improvements in programme quality are greatest.

Insist on full disclosure in evaluation reports

The choice of methods and evaluation parameters used, methodological drawbacks, and areas of subjective judgement should be described in full. Evaluation studies should make clear what can and cannot be quantified. As outcomes can be expressed in different ways using different variables, choices about how results are expressed should be explicitly stated.

Perform evaluation using independent experts as well as in-house staff

If in-house staff administer surveys of programme beneficiaries, especially of clients they have worked with directly, there is a heightened risk of response bias. Programmes should also be evaluated by, or in collaboration with, independent external experts, possibly from an Audit Office. Ideally, the body that implements the evaluation would work with programme managers but would not be dependent on continued contracts from the sponsor of the programme. However, an issue here is the independence of in-house staff who take part in evaluations. Staff need to be safe from any form of sanction in cases where evaluation findings prove negative.

Encourage evaluators to consider the implications of evaluation findings for various possible courses of policy

For instance, an evaluation with a future-oriented perspective might highlight the fact that diminishing returns would likely arise in a given programme were it be expanded.

Invest in the collection of appropriate local-area data

Administrative data is often unavailable for small geographic areas. And different parts of government sometimes collect data across geographic units of different scales. There is a need for statistical systems that provide easily accessible data for small geographic units that remain stable over time. It is essential to collect baseline data that are relevant to the goals of policy, can be obtained across target and comparator localities and firms, and that can also be tracked over time. Collecting data at the point of programme delivery can be much less costly than *ex post* data collection. Especially in pilot phases of programme development, programme assistance might be directed to areas for which detailed statistical information is available. Some countries are beginning to address the local area data problem. For example, as described by Robson in this volume, the UK government has invested in establishing a National Neighbourhood Statistics Database (*neighbourhood.statistics.gov.uk*) using local authority wards as a reporting unit.

For central governments

Be strategic in the design of support for local economic developmentin order to facilitate identification of control-group localities

Locations that are home to particular policy initiatives sometimes have few comparator localities if the policy in question is highly inclusive. For example, if programmes are implemented in all locations that have some cut-off rate of unemployment, the only places left to compare against will be those that probably have important dissimilar features.

Aim to achieve coherent engagement among different tiers of government

The involvement of different tiers of government in jointly funding, designing and implementing evaluations appears rare. In the United States, for instance, genuine intergovernmental evaluations have been infrequent (Morra, 1997). The United States' federal government has had a lead role, although states have become increasingly active in evaluation. In France, in evaluating planning contracts between the central government and the regions, each region developed its own methods and procedures. The need to co-ordinate evaluations was articulated in a Prime Ministerial circular in 2000 (Heddebaut, 2000).

Consider providing financial support for evaluations done locally

This and other chapters in this book describe why local bodies are likely to under-invest in high-quality evaluations. Demands that some types of programme achieve financial autonomy might also discourage spending on evaluation. Consequently, there is a case for central governments to provide financial or other support for evaluations done locally. Such support could also include the provision of training for evaluators.

Develop clear evaluation standards and guidelines

Public authorities should seek to develop evaluation standards and guidelines. Central authorities in some countries already do this (for example, the General Accounting Office in the United States). Among other benefits, evaluation standards and guidelines could help local governments to state the methodological norms expected of subcontracted companies that perform evaluations. Indeed, adherence to evaluation standards could be made a condition for programme support and contracts. Aside from upgrading the quality of evaluation, greater uniformity in evaluation practice could also help redress the current situation in which programme outcomes are often not comparable because of the use of opaque and/or non-standard evaluation methods. Indicative standards could even apply to survey design.

For local governments

Local authorities should take an active role in developing case studies and surveys of programme participants.

Surveys are likely to be most reliable when they address programmes that provide real services rather than financial assistance. Surveys can be of value in helping to shed light on why a programme has or has not worked and what might be done to improve effectiveness through alternative designs and modes of implementation. When doing surveys, questions need to be

formulated in a highly specific manner. Questions must address the counterfactual and focus on key features of programme design. Surveys should be based on samples large enough to draw valid inferences from.

Seek to partner with academic bodies

Evaluation is a technical and specialized discipline. Often, local authorities do not possess the necessary in-house expertise. Knowledge of economics, statistical theory and data characteristics are critical to rigorous evaluation. One way to bridge capability gaps would be to form evaluation partnerships with – and/or generally draw on expertise available in – academic institutions. Academics may also be willing to analyse data at low cost if the results are likely to be of interest to the academic community. Publication of the relevant research should thus also be encouraged. Being able to draw on academic expertise to ensure appropriate terms of reference can be important when subnational governments subcontract evaluative research to private consultancy firms. Academic expertise can also help in comparing the relative quality of different evaluation studies, given that policy conclusions might be based on work that is weak from a methodological, theoretical or data-quality perspective. Local authorities might also create evaluation partnerships through, or with, national associations of local governments. Such inter-governmental partnerships might aim to disseminate evaluation findings among subnational bodies and secure competent technical advice.

Notes

1. Policymakers face choices in allocating public resources to different uses. In a textbook world local policymakers would allocate resources to different programmes based on knowledge of the *marginal* cost of achieving given common objectives for different programme types. For instance, if employment creation were the principal goal of local development policy, an economically efficient resource allocation could be achieved if policymakers had information on the *marginal* costs of creating a job through such programmes as investment attraction, enterprise start-up grants, training, etc. An efficient allocation would obtain when the marginal costs of job creation were the same across programmes (Storey, 1990). In practice, however, most evaluations provide information on the *average* cost of employment creation. In addition, the marginal cost of job creation will vary over time depending, for instance, on the scale and duration of a programme and the key features of the local labour market. The textbook ideal, then, would necessitate a constant cycle of complex evaluation across many programme types, and is effectively unattainable.
2. Often, partnerships are treated as a good in themselves. Discussion of possible opportunity costs or displacement effects (see footnote 4) from working through partnerships appears to be rare.
3. Bartik also notes that it is unclear whether the returns to targeting resources for area development are higher in more or less distressed areas. However, it is

unlikely that there could be a single response to this question. In practice, whether a particular programme will have a greater or lesser impact in more or less distressed areas depends on the programme type. There are of course many different types of area-development programme in use. Some programme types might be more effective on some evaluation criteria than on others, and variation in this respect might be sensitive to the degree of distress encountered. For instance, entrepreneurship support schemes in wealthier areas will often cater to larger firms and better qualified entrepreneurs. They may therefore be associated with greater enterprise longevity and possibly larger overall multiplier effects. Enterprise displacement might also be lower in wealthier than poorer areas. So entrepreneurship support may work well on these evaluation parameters in wealthier places. However, because better qualified entrepreneurs are more likely to have entered entrepreneurship in the absence of programme support, the deadweight associated with such schemes might be higher in wealthier localities (OECD, 2003).

4. Deadweight refers to the extent to which outcomes would have occurred in the absence of the policy intervention. For instance, in the absence of a formal partnership initiative, other institutions or individuals may have come together anyway to achieve similar objectives. Displacement essentially refers to a loss of output, employment or opportunity stemming from a programme or policy. Displacement takes many forms across a range of markets. For example, interventions might use resources – such as skills or land – that become unavailable for alternative uses. Or the price of these resources may rise. And from a community development perspective, programmes to encourage localised networking could conceivably lead to a reduction in networking activity across a wider area [Armstrong *et al.* (2002)]. Interventions financed from general taxation will also have a displacement effect on private consumption or investment elsewhere in the economy (although these effects may be offset by supply side improvements to the economy). If displacement occurs across rather than within localities, then its significance from an area-development perspective will also depend on whether displacement occurs in areas that are more or less affluent than the target locality. Substitution effects are often described in connection with labour market policy. For example, substitution occurs when trainees fill vacancies that would otherwise have been filled by non-trainees.

5. However, it is not infrequent that guidelines recommend actions to evaluators that are almost inoperable. For instance, HM Treasury advocates that "Evaluation studies should always provide information on displacement on a local and national (UK) basis". This is too demanding.

6. In this connection, Malan (2002) reports on a major *ex post* evaluation of the European Commission's 1989-1993 Objective 2 programmes. He notes that both the European Regional Development Fund and the European Social Fund were eager to estimate synergies occurring between co-located programmes. Only three of the 25 evaluators involved attempted to apply the Commission's evaluation methodology (MEANS) to this end, judging the task too complex and time-consuming.

7. Barkley (2003), for instance, documents how State-level policymakers in the United States have failed to allow the time necessary to judge the effectiveness of innovative programmes of equity finance in rural areas. Such programmes can involve equity funds that have investment cycles of five years or more. Policymakers have often introduced new generations of scheme prior to receiving a full assessment of existing programmes.

8. The term "social enterprise" usually refers to firms that aim to achieve social objectives using resources from a variety of public and private sources. The combination of public and other funding with income earned from market transactions is the key trait distinguishing social enterprises from traditional non-profit organisations.

References

ARMSTRONG, H.W., KEHRER, B., WELLS, P. and WOOD, A.M. (2002), "The Evaluation of Community Economic Development Initiatives", *Urban Studies*, Vol. 39, No. 3, pp. 457-481.

BARKLEY, D.L. (2003), "Policy Options for Equity Financing for Rural Entrepreneurs", in *Main Streets of Tomorrow: Growing and Financing Rural Entrepreneurs*. Center for the Study of Rural America, Federal Reserve Bank of Kansas City, October, 2003, pp. 107-126.

BARTIK, T.J. and BINGHAM, R.D. (1997), "Can Economic Development Programs be Evaluated?" in Bingham, R.D. and Mier, R. (editors), *Dilemmas of Urban Economic Development*, Sage Publications.

BOARNET, M.G. (2001), "Enterprise Zones and Job Creation: Linking Evaluation and Practice", *Economic Development Quarterly*, Vol. 15, No. 3, August 2001, pp. 242-254.

BONDONIO, D. and ENGBERG, J. (2000), "Enterprise Zones and Local Employment: Evidence from the States' Programs", *Regional Science and Urban Economics*, 30, pp. 519-549.

CANADA (2000), *Study of the Evaluation Function in the Federal Government*, Treasury Board of Canada Secretariat, Ottawa.

COWLING, M. and HAYWARD, R. (2000), *Out of Unemployment*, Research Centre for Industrial Strategy, United Kingdom.

DETR, DEPARTMENT OF THE ENVIRONMENT, TRANSPORT AND THE REGIONS (2001), *A Review of the Evidence Base for Regeneration Policy and Practice*, London.

EISINGER, P. (1995), "State Economic Development in the 1990s: Politics and Policy Learning", *Economic Development Quarterly*, Vol. 9, pp. 146-158.

FOLEY, P. (1992), "Local Economic Policy and Job Creation: A Review of Evaluation Studies", *Urban Studies*, Vol. 29, No. 3 and 4, pp. 557-598.

HEDDEBAUT, O. (2000), "Decentralization and Evaluation: the Problems Caused by the Various Levels of Decision and the Length of the Observation Period", paper prepared for the Fourth European Evaluation Society Conference, Lausanne, October 12-14.

HM TREASURY (1995), *A Framework for the Evaluation of Regeneration Projects and Programmes*, London.

MALAN, J. (2002), "Translating Theory into Practice: Lessons from the *ex post* Evaluation of the 1989-1993 Objective 2 Programmes", Paper presented at the European Evaluation Society 2002 Conference, Seville, Spain, October 10-12.

MEIER, J.K, and GILL, J. (2000), *What Works: A New Approach to Program and Policy Analysis*, Westview Press, Oxford.

METCALF, H., CROWLEY, T.V., ANDERSON, T. and BAINTON, C. (2000), *From Unemployment to Self-Employment: The Role of Micro-Finance*, International Labour Office, London.

MORRA, L.G. (1997), "Evaluation in the United States: Cooperative but not Intergovernmental", in Rieper and Toulemonde: *Politics and Practices of Intergovernmental Evaluation*, Transaction Publishers.

OECD (1999), *Improving Evaluation Practices*, OECD, Paris.

OECD (2003), *Entrepreneurship and Local Economic Development: Programme and Policy Recommendations*, OECD, Paris.

DEPARTMENT FOR TRANSPORT, LOCAL GOVERNMENT AND THE REGIONS (2002), "Lessons and Evaluation Evidence from Ten Single Regeneration Budget Case Studies", Mid-term Report, London.

RIEPER, O. and TOULEMONDE, J. (1997), *Politics and Practices of Intergovernmental Evaluation*, Transaction Publishers.

SCHMID, G., O'REILLY, J. and SCHÖMANN, K. (1996), "Theory and Methodology of Labour Market Policy and Evaluation: An Introduction", in Schmid, O'Reilly and Schömann (eds.), *International Handbook of Labour Market Policy and Evaluation,* Edward Elgar, Cheltenham.

SHAPIRA, P. (2003), "Evaluating Manufacturing Extension Services in the United States: Experiences and Insights", in Shapira, P. and Kuhlmann, S., *Learning from Science and Technology Policy Evaluation*, Edward Elgar, Northampton, Massachusetts and Cheltenham, United Kingdom.

STATE SERVICES COMMISSION (1999), "Looping the Loop: Evaluating Outcomes and Other Risky Feats", Occasional Paper No. 7, Wellington, New Zealand, June.

STOREY, D.J. (1990), "Evaluation of Policies and Measures to Create Local Employment", *Urban Studies*, Vol. 27, No. 5, pp. 669-684.

THE AUDIT COMMISSION FOR LOCAL AUTHORITIES AND THE NATIONAL HEALTH SERVICE (1999), *A Life's Work: Local Authorities, Economic Development and Economic Regeneration*, September, United Kingdom.

WOOLCOCK, M.J.V. (1999), "Learning from Failures in Micro-finance", *The American Journal of Economics and Sociology*, 58 (1), pp. 17-42.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Progrmmes and Policies
© OECD 2004

Chapter 2

Policy Learning through Evaluation: Challenges and Opportunities

by
Ging Wong,
Canadian Heritage and University of Alberta

Context

Imagine a new delegate going to his first meeting of the OECD Local Economic and Employment Development (LEED) Committee. While eager to share his country's experiences, he is also excited at the prospect of learning from others on a wide range of economic and social innovations through locally-based initiatives. He is not disappointed, as LEED offers solid analytical and data resources to serve Members' discussion, working groups to compare experience and seek answers to thematic policy issues and, above all, hundreds of practical case studies that document important successes and failures. But he is frustrated by the problem of distilling the essence from this very rich source of information – key lessons that potentially can improve the quality of public policies and operations.

That delegate was me seven years ago. Since then, I and others have worked closely with the Chairs of the LEED Committee and the LEED Secretariat to develop a more systematic and rigorous assessment of current practices. In short, we are promoting an evaluation culture, not as an end in itself but as a valuable information tool. The conference, on which this book is based, represents the launch of a new enterprise and a beginning of a new dialogue.

Introduction

Let me present some initial thoughts on policy learning through evaluation.

If quality, relevance and timeliness are the hallmarks of good evaluation, then we are well served by the work of the European Evaluation Society. I was particularly struck by the Society's 2002 conference theme "Three movements in Contemporary Evaluation: Learning, Theory and Evidence". One of the aims of the Vienna Conference was to explore the essential relationships between these elements and their implications for local economic and employment development public policies. I will concentrate on the policy learning potential from evaluation.

Let me start by quoting the European Evaluation Society conference brochure:

"Learning is widely seen as an overarching purpose of evaluation. The learning movement in evaluation began with concerns that policy

makers should learn from evaluation. But in a world where the State no longer tries to do everything and is often in partnership with 'civil society', learning through evaluation is not only the preserve of policy makers. Communities, associations and citizens also use evaluation to learn: to improve their understanding of available choices and to help develop new forms of 'inclusive' consensus. A second impetus that is re-shaping learning as a movement in evaluation today comes from the fields of knowledge management, organisational learning and organisational memory. Evaluators are becoming increasingly concerned with how we hold on to what we have learned: that past evaluation findings are not lost and can be accessed and retrieved when needed in the future."

This quotation elegantly describes the power of evaluation information, which presents evidence of performance and impact that should properly be interpreted within the appropriate policy context and theoretical framework, for various target audiences concerned with whether policies and interventions are working or not. It also suggests several purposes for evaluation, reasons why governments might want to invest in developing evaluation systems, namely to demonstrate public accountability, promote democratic processes, and establish a substantial knowledge base for research and policy development. These are all good things to support.

However, policy learning from evaluation also encounters many challenges as well as opportunities. To explore these, this paper will proceed from a discussion of the present scope and uses of evaluation activities to the special needs of local development and employment programs. My comments are shaped by my work as a former Director of the largest evaluation group in the Canadian government and my interaction with the international community through the World Bank, OECD and bilateral policy and technical exchanges between Canada and other countries.

Scope and uses of evaluation activities

In taking stock of the scope and uses of evaluation, my starting point is to say that evaluation means many things to many people, especially in an international forum of multiple audiences. Some twenty years ago, according to Glass and Ellett (1980), "evaluation – more than any science – is what people say it is, and people currently are saying it is many different things".[1] While there continues to be lively debates about the nature and uses of evaluation – for example, the benefits and uses of evaluation by client group, and how evaluations can be better designed and implemented to ensure use – changes and development in evaluation practices have given us a better understanding of what it is and what it is not.

Related but not the same

First, evaluations are not audits. There is still a general confusion between these activities, in part because evaluation is a commonplace term for any type of assessment or where other forms of applied research such as action research and strategic analysis are often considered to be substitutes. It is also partly because some national audit offices have broadened their remit from traditional financial audit of public expenditures with due regard to efficiency and economy to "management control systems" and "value for money" (VFM) audits on the effectiveness of resource management. Indeed, in Canada, the 2001 budget of the Office of the Auditor General (OAG) shows 60 per cent of the budget goes to VFM or what is now called "performance auditing" on how well policies and programs have been implemented, with 33 per cent on "financial statement audit activities" (Sutherland, 2002). To some commentators, this represented a strong growth in the regulation of government activities. This *audit explosion*, as Michael Power has called it, has created a new accountability that takes the form of detailed control with sharp teeth. As recently described by Onora O'Neill in her 2002 BBC Reith Lectures on *A Question of Trust*:

> "The new accountability culture aims at ever more perfect administrative control of institutional and professional life… The idea of audit has been exported from its original financial context to cover ever more detailed scrutiny of non-financial processes and systems. Performance indicators are used to measure adequate and inadequate performance with supposed precision."

Second, evaluations are complementary to but different from results-based management systems that are increasingly part of public sector regulation (Davies, 1999). Much of the motivation for performance management is attributed to the growth of the New Public Management (NPM) of the 1980s, with its call for decentralizing program and service responsibilities while ensuring that public institutions are accountable for program results. Results-based accountability requires that organisations demonstrate how public monies will be spent: by articulating what programs are to achieve in terms of outcomes, putting into place indicators and data collection instruments to measure whether or not outcomes have been achieved, setting performance standards or benchmarks to assess how programs are progressing, and periodic analysis for internal decision making and public reporting (Horsch, 1996). In short, results-based accountability systems answer three questions: What is the program trying to achieve? What program progress is made? Have desired results been achieved? Program evaluation answers different, yet equally important, questions: Why is the program succeeding or failing? What unintended results have resulted? What

changes might be necessary to improve effectiveness? Performance management then asks the "what" questions whereas evaluation asks the "why" and "how" questions.

In the most recent statement of Canadian government evaluation policy (Treasury Board Secretariat, 2001):

- Managing for results is the prime responsibility of public service managers… are expected to define anticipated results, continually focus attention towards results achievement, measure performance regularly and objectively, and learn and adjust to improve efficiency and effectiveness…
- Evaluation can support managers' efforts to track and report on actual performance and help decision-makers objectively assess program or policy results. This distinguishes evaluation from internal audit – a function that provides assurances on a department or agency's risk management strategy, management control framework and information, both financial and non-financial, used for decision-making and reporting.

Accordingly, government departments are instructed to "embed the discipline of evaluation into the lifecycle management of policies, programs and initiatives to:

- develop results-based management and accountability frameworks for new or renewed policies, programs and initiatives;
- establish ongoing performance monitoring and performance measurement practices;
- evaluate issues related to the early implementation and administration of the policy, program or initiative, including those that are delivered through partnership arrangements (formative or mid-term evaluation); and
- evaluate issues related to relevance, results and cost-effectiveness."

Program evaluation then provides effectiveness information required for a PPBS-style reporting system (with its central controls) and, when linked to results-based management (with its emphasis on letting managers manage), create accountability systems which are a powerful management tool.

Benefits and uses of evaluation

> *Being efficient was about accounting for expenditure; being effective was about achieving meaningful results (Parata, 1998).*

This reminds us of that wicked episode in that wonderful BBC television series *Yes Minister* where the most efficient hospital in the Minister's portfolio was not only efficient but had no patients at all. When the Minister questioned whether this made any sense, the reply was "patients would get in the way of their efficiency performance measures". The point here is not to disparage efficiency but to remember the importance of meaningful results.

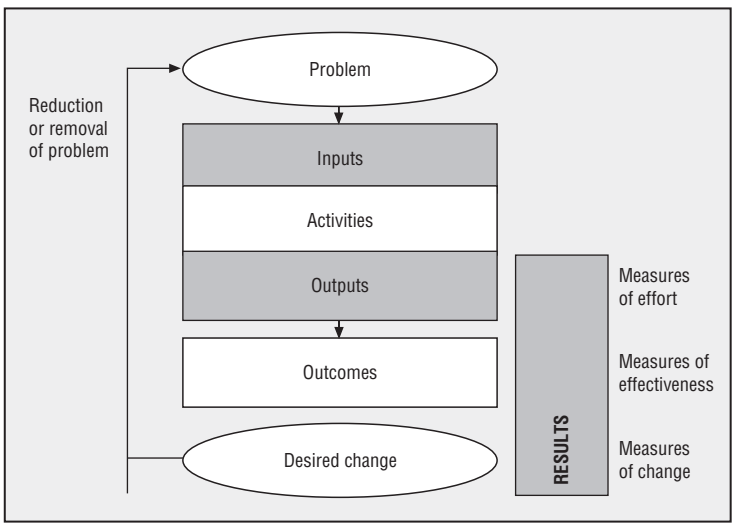
Evaluation, in the Canadian context, is essentially concerned with effectiveness, which poses the counterfactual question of what would have happened if the program did not exist. One could use observational or correlation methods for demonstrating whether desired effects occurred, and quasi-experimental and experimental designs for determining whether the observed effects can reasonably be attributed to the intervention and not to other sources. Four key issue categories relate to program rationale, continued relevance, impacts and effects, and alternatives.

In practical terms, an evaluation tries to determine as a result of a policy, program or initiative (Fowler, 2002):

- What has changed? Identification of difference.
- What caused the change? Attribution.
- Was it what was intended? Judgment.
- What next? Learning and adjustment.

Figure 2.1 shows a typical logic model used to map the causal processes of an intervention and test them as part of an evaluation.

Figure 2.1. **Typical intervention evaluation logic model**



Changing landscape

We should note, however, that the scope of evaluation ranges from a singular purpose to assess the value of a policy or program consistent with the desired outcomes to a broader approach with multiple benefits.[2] For example,

in 1980 Scriven wrote that "Evaluation is what it is, the determination of merit or worth, and what it is used for is another matter… Bad is bad and good is good and it is the job of the evaluator to decide which is which". This is countered, however, by Weiss who saw "the purpose of evaluation research is to measure the effects of a program against the goals set out to accomplish, as a means of contributing to subsequent decision-making about the program and improving future programming". Later definitions of evaluation consistently reflect a broader approach, with greater emphasis on the various benefits associated with evaluation. Patton (1997) provides perhaps the quintessential definition:

> "Program evaluation is the systematic collection of information about the activities, characteristics and outcomes of programs to make judgments about the program, improve program effectiveness, and/or inform decisions about future programming."

In other words, evaluations can be used for making overall judgments, facilitating improvements and generating knowledge. In her review of the literature, Chelimsky (1997) agrees that almost all other specified benefits are subcategories of these three conceptual areas: accountability, development and knowledge. Let us briefly look at each of these in turn.

Accountability and decision-making

Evaluation can support accountability for program performance and spending. Modern accountability systems and accountability evaluation are rooted in the history of government-wide expenditure budgeting. In 1971, the Planning, Programming and Budgeting System (PPBS) was brought into Canada from the US and embedded into the federal government reporting format for the annual Estimates of Expenditure, despite the fact that the US government had dropped PPBS two years earlier. The American experience since 1965 had shown that it was very difficult to generate the objective information needed and that subjective judgments flowed into the vacuum created. Ultimately, PPBS centralized power in the administration at the expense of Congress, and so Congress got rid of it.

In Canada in 2002, the legacy of PPBS and its various mutations is the same intellectual problem it presented in 1971 – how to capture program results to satisfy central agency requirements for highly-aggregated statements of program accomplishment that are demonstrably true and verifiable (Sutherland, 2002). Program evaluation was created in 1977 as a mandated function in all federal government ministries to address the analytical burden of determining program impacts on a mandatory review cycle ranging from 3-5 years. In the accountability perspective, evaluation is "faced with cause-and-effect questions requiring methods able to link findings to interventions as closely and conclusively as possible" (Chelimsky, 1997).

A similar European story is documented by Jacques Toulemonde (2000, 351):

> "It is well known that evaluation was born in the United States along with Planning-Programming-Budgeting-System (PPBS). It was imported in the 1970s into most northern European countries where agencies, units or commissions were created to carry out policy analysis. These institutions dealt with *ex ante* and *ex post* evaluation mission. They had an inter-ministerial scope and they clearly aimed at introducing some scientific rationality in the budgetary process. Countries such as the United Kingdom, the Netherlands and France imported the US model with enthusiasm…"

All countries that have known this type of evaluation have now gone beyond it or abandoned it. The most abrupt and complete change took place in France where PPBS was done away with in the early 1980s. However, "the baby was thrown out with the bath water", to such an extent that evaluation lost all significant support at government level for nearly ten years.

In other countries the transition was far more progressive. For example, in the Netherlands the Finance Ministry made a smooth change from PPBS-based evaluation to a far more pragmatic system. In the UK the Policy Analysis Commission set up in the 1970s was dissolved when the Thatcher government re-launched evaluation with the "value for money" slogan.

Since the mid-1980s, the diffusion of evaluation on a European scale was boosted by professional networks in international policy research and development aid. The OECD created vigorous evaluation workgroups as part of its Development Assistance Committee and PUMA. The World Bank and other international institutions such as the ILO also helped to spread the culture of evaluation, either as a requirement for aid funding to developing countries or as a project management tool. A more decisive role in capacity-building was played by the European Union through its regulations concerning Structural Funds. Rules adopted in July 1993 for the period 1995-1999 made *ex ante* evaluation a pre-requisite for funding, and monitoring and *ex post* evaluation were made mandatory for regional and national authorities. The number of evaluations related to Structural Funds was multipled by five to six fold, compared to the previous five-year period. In addition to these external push factors can be added internal driving forces peculiar to certain countries: pressure from parliament in Germany and Scandinavia, initiatives by the national audit offices in the Netherlands, Sweden and the UK, and pressure from finance ministries, as in Italy (Toulemonde, 2000, 352-353).

At the same time, the observation is made that evaluation demand is developing faster at European and national level than at regional and local level (Leeuw, Toulemonde and Brouwers, 1999):

> "While almost all European policies are now subject to periodic evaluation, a figure of 10 per cent was mentioned for the proportion of

regional policies that are subject to periodic evaluations in countries like the UK, Denmark, the Netherlands and France."

Clearly, this limited penetration of evaluation of policies at the sub-national level presents a challenge to the OECD Local Economic and Employment Development (LEED) Program.

Development

Evaluations can also facilitate policy and program development by:

- Improving program design based upon past evaluation findings and the clear articulation of realistic, attainable objectives.
- Improving program implementation through process evaluations.
- Improving program cost-effectiveness by investigating alternative ways of meeting objectives at the lowest cost.
- Supporting effective management practices by validating indicators and performance targets, identification of effective practices.
- Build analytical and measurement capacity.

Formative methods typically used in developmental evaluations include: case studies, research synthesis, internal self-evaluations, performance measurement and monitoring. Such process evaluation activities do not usually examine results, which is usually left until the program is well established.

Knowledge and skills

At a minimum, evaluation can build knowledge about existing/potential needs, and about programming that address those needs, by:

- Increasing knowledge of needs and problems.
- Increasing knowledge of effective practices and programs.
- Increasing knowledge of programming.

Evaluation can develop capacity for better program design, assessment and improvement by:

- Learning to think more critically about programs.
- Improving attitudes toward evaluation.
- Developing capacity to understand, use and conduct evaluation.

Unlike judgment-oriented and development-oriented evaluations where decisions may follow from the findings, knowledge-oriented evaluation is used to influence thinking and deepen understanding by increasing knowledge. This conceptual or enlightened rather than instrumental use of evaluation findings "can be as specific as clarifying a program's model, testing theory,

distinguishing types of interventions, figuring out how to measure outcomes, generating lessons learned, and/or elaborating policy options" (Patton, 2001: 332). Local program people gain new ideas and insights from evaluations that can change their understanding of what the program is and does.

Utility and humility

By now there is a large literature about the use and non-use of evaluation. This developed in response to a crisis in confidence in the 1970s when studies found that evaluations were not always routinely used as the central input into policy decisions. Subsequently, a great deal of attention has focused upon the multiple uses of evaluation that are integral to its purpose and design. The general categories of accountability, development and knowledge are suggested as a result of research on "achieved (not hoped for) use" from certain types of evaluations (Chelimsky, 1999). Certainly, it is my experience that prospective, formative and summative evaluations all have very different uses and relevance at different times of the policy development cycle. Prospective evaluations provide a synthesis of existing evidence through meta-analyses or exploratory analysis of a synthetic data base. The results can inform the initial stages of policy analysis where problem identification and alternative solutions are examined. Formative evaluations of new program implementation inform decision-makers whether the appropriate ingredients are put in place effectively while summative evaluations measure and account for the results of the new program. These are a only small sample of evaluation approaches, but they do illustrate that an evaluation cannot be all things to all people.

Yet there are grumblings that the focus on differential use somehow trivializes evaluation. Thus we are often reminded of the higher purpose of evaluation.

> "Evaluation is learning from experience to improve future work. Therefore all evaluations should aim at the same objective or purpose as the projects, programs and policies evaluated: the creation of *sustainable benefits for target groups*. They cannot do so if they are not used. Therefore, all evaluations must aim at usefulness." (Eggers, 1999: 93)

> "The goal of evaluation is not use, the goal is social betterment. Use, then, while not the ultimate goal is a means by which evaluation achieves social betterment." (Henry, 2000: 3)

At the broadest level, as Chelimsky points out, these may indeed express the ultimate purpose of evaluation. Over the long-term, the local benefits of evaluation will contribute to more effective social programming, financial savings and improved well-being. Yet, in the short term, there are specific and immediate uses, somewhere lower in the food chain.

Furthermore, as Eggers admits, evaluations "are not necessarily used even if they are usable. The evaluation must lead the horse to the water even if cannot make him drink". This recognizes that evaluations, even with compelling empirical evidence, represent only one policy consideration and compete against other political, and public opinion inputs. It is also true that increasingly shorter policy development cycles discount rigorous evaluations that cannot be delivered within established budgets and timeframes, relying instead upon more impressionistic evaluations. Despite these limitations, evaluation is one of the few sources of credible evidence on policy impacts and effectiveness. To increase its overall utility, evaluation should be sustained as an important activity in the long-term and a repository of knowledge that could be revisited for particular enquiries.

Above and beyond that, there is the real problem of communicating the results of evaluations in plain language and in the proper context in order that policy makers take note. Evaluators are advised by Weiss (1998) "to work hard at communicating their findings… (but) that they should not hold out unrealistic expectations for use." Patton (1997), on the other hand, argues that participatory strategies with program stakeholders to enlist their knowledge and commitment can increase the use of evaluation findings.

Challenges and opportunities

I am of the view that evaluation and subsequent policy learning takes place in an organized context. For local economic and employment development policies and programs, the organized context include the following groups of potential users – program sponsors (*e.g.* government, non-profit organisation, foundations), program staff and operators, program clients, and civil society or that part of the public that is active in communal and associational life. These are the local partners who experiment and innovate in resolving their economic and social problems. The challenge is to create and seize opportunities to develop participatory evaluation strategies for multiple stakeholders. Evaluation's contribution here is to provide trustworthy and accurate effectiveness information to inform decision-making. But it can also help local partners learn to weigh evidence and think in an evaluative way, something which may have a more enduring impact than specific findings.

The task at hand is a considerable one if the estimate is accurate that only 10 per cent of regional and local policies are presently evaluated. Yet local-level case evaluations can illuminate and provide context to more aggregate indicators. And synthesis evaluations and meta-analyses that combine community-based evaluation findings "to generate more system-wide patterns and lessons is an increasingly important way of integrating decentralized evaluations for use by policymakers at the central system level"

(Patton, 2000: 15). However, if evaluations based upon individual-level data were adopted for communities, there is the methodological challenge of converting measures of effectiveness to impact findings for a locality.

We should also remember that evaluation strategies must be appropriate to the innovative character of local development programs and initiatives. Prospective rather than retrospective evaluations may be better suited here.

Further, evaluations must also recognize the horizontal nature of local development policies and programs which often cuts across jurisdictions. Social cohesion, social capital and social innovation are difficult phenomena to collect data on systematically and to quantify. A multi-disciplinary approach is warranted.

Despite these challenges, the potential for policy learning in this domain is huge and the insights gained through evaluation and exchange make this a worthy endeavour.

Notes

1. Cited in Heinecke, Blasi, Milman and Washington, 1999, p. 1.
2. The following quotations are cited in McGuire (2002).

References

CANADA Office of the Auditor General of Canada. Annual Reports of the Auditor General for the Fiscal Years, 1982-83 (Chapter 3), 1992-93 (Chapters 8, 9, 10), 1995-96 (Chapter 3), 2002-02 (Chapter 6), (*www.oag-bvg.gc.ca*).

CANADA Treasury Board. Evaluation Policy, Ottawa, February 2001 (*www.tbs-sct.gc.ca/pubs_pol/dcgpubs/tbm_161/ep-pe_e.html*).

CHELIMSKY, Eleanor, "Thoughts for a New Evaluation Society", Evaluation, Vol. 3 (1), pp. 97-118.

DAVIES, Ian C. "Evaluation and Performance Management in Government", Evaluation, Vol. 5 (2): 150-159.

EGGERS, Helmut W. and Eleanor CHELIMSKY "Purposes and Use: What Can We Expect?" Evaluation, Vol. 5 (1): 92-96.

FOWLER, Alan, Measuring Non-tangible Outcomes: The Challenge for Impact Assessment, Australia, April 2002 (*www.mande.co.uk/docs/fowler%20presentation.pdf*).

HEINECKE, W.F., Laura BLASI, Natalie MILMAN and Lisa WASHINGTON, New Directions in the Evaluation of the Effectiveness of Educational Technology, The Secretary's Conference on Educational Technology, 1999 (*www.ed.gov/technology/techconf/1999/whitepapers/paper8.html*).

HENRY, Gary T., "Using Evaluation Findings for Policy: A Realist Perspective", paper presented at the 4th European Evaluation Society (ESS) Conference, Lausanne, 12-14 October 2000.

HORSCH, Karen, "Results-based Accountability Systems: Opportunities and Challenges", Harvard Family Research Project, The Evaluation Exchange, Vol. II, No. 1, Winter 1996 (*http://gseweb.harvard.edu/~hfrp/eval/issue3/horsch.html*).

KVTASTEIN, Olav A., "Extracting Lessons from Central Debates in order to Improve Policy Implication of Evaluations", paper submitted for the 4th EES Conference, Lausanne, 12-14 October 2000.

MACKAY, Keith (ed.), Public Sector Performance – The Critical Role of Evaluation. World Bank, Washington, DC 1998.

MCGUIRE, Martha, Literature Review on the Benefits, Outputs, Processes and Knowledge Elements of Evaluation. Canadian Evaluation Society Project in Support of Advocacy and Professional Development. Ottawa, October 2002.

O'NEILL, Onora, BBC Reith Lectures 2002 – A Question of Trust (*www.bbc.co.uk/radio4/reith2002*).

OECD, Evaluation Feedback for Effective Learning and Accountability, Development Assistance Committee (DAC) Working Party on Aid Evaluation, Paris, 2001.

OECD, Improving Evaluation Practices: Best Practice Guidelines for Evaluation and Background Paper, PUMA, Paris, 1999.

PARATA, Hekia, "The Outputs/Outcomes Nexus: the relative responsibilities between Ministers and the Public Service", in Lifting the Game from Outputs to Outcomes: Proceedings, Public Service Senior Management Conference, State Services Commission, Wellington, January 1998.

PATTON, Michael Q., Utilization-focused Evaluation: The New Century Text. Sage. California: Thousand Oaks, 1997.

PATTON, Michael Q., "A Vision of Evaluation that Strengthens Democracy", Keynote to the 4th EES conference, Lausanne, 2000.

PATTON, Michael Q., "Evaluation, Knowledge Management, Best Practices, and High Quality Lessons Learned", American Journal of Evaluation, 2001, Vol. 3 (1): 329-336.

SUTHERLAND, S.L., The Office of the Auditor General of Canada: Government in Exile? Working Paper 31, School of Policy Studies, Queen's University, Kingston, Ontario, September 2002.

TOULEMONDE, Jacques, "Evaluation Culture(s) in Europe: Differences and Convergence between National Practices", Quarterly Journal of Economic Research, 3/2000, S.350-357 (*www.diw.de/english/publikationen/vierteljahrshefte/jahrgang00/content_3.html*).

UNITED STATES, Case Study Evaluations. General Accounting Office, Washington, DC November 1990.

WEISS, Carol H., "Have We Learned Anything New About the Use of Evaluation?" American Journal of Evaluation. Winter 1998, Vol. 19 (1).

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 3

Evaluation: Evidence for Public Policy

by
Robert Walker,
Professor of Social Policy,
University of Nottingham, United Kingdom

This contribution provides an overview of the political, institutional and methodological challenges that confront public policy evaluation with a view to stimulating a constructive, collaborative response. It begins by specifying the different questions asked in policy evaluation and links these to techniques for answering them. The second section focuses on the threats to effective policy evaluation that arise from the policy process, the nature of policy and the marketplace for evaluation. The final section grapples with the challenges of building and nurturing a vibrant and sustainable culture of evaluation. Extensive use is made of case-studies, both of specific policy evaluations and of the United Kingdom which, since the election of a *New* Labour government in 1997, has sought to introduce a culture of policy evaluation into the heart of central government policy making.

"Will this policy work?" "What kind of policies have worked in the past?" If questions such as these could be answered satisfactorily, much of the risk would be taken out of politics and policy making turned into a science.

Reality falls far short of such an aspiration. Not only are these kinds of question rarely asked of social science, social scientists seem unable to answer them with much sense of security. After a great deal of effort, robust answers may be provided to tightly prescribed questions that limit generalisation, while the big strategic questions generate imprecise answers hedged by qualifications. Consequently, evaluation evidence is likely to remain just one of the many sources of information deployed by policymakers and politics will continue to be a risky enterprise. To the extent that this makes politics a degree less boring, it may actually be good for the health of democracy.

Nevertheless, reducing the risk of introducing poor policies or rejecting good ones by even a small amount could prevent billions of dollars, euros or pounds from being wasted and add greatly to the sum of human well-being. It is therefore incumbent on the policy community to seek to improve the quality, quantity and use of evaluative evidence. This will necessitate a close partnership between those seeking evaluative evidence – who need to ask for it sufficiently early, help fund its generation and be prepared to act on both welcome and unwelcome findings – and producers who have to engage with the relevancies of policymakers while protecting and advancing standards of scientific enquiry. This, in turn, requires institutions to promote, foster and facilitate the partnership by encouraging productive dialogue and ensuring that the diverse rewards are appropriately shared.

This overview of the issues begins by specifying the different questions asked in policy evaluation and links these to techniques for answering them. The second section focuses on the threats to effective policy evaluation that are born of the policy process, the nature of policy and the marketplace for evaluation. The final section grapples with the challenges of building and nurturing a vibrant and sustainable culture of evaluation. Extensive use is made of case-studies, both of specific policy evaluations and of the United Kingdom which, since the election of a *New* Labour government in 1997, has sought to introduce a culture of policy evaluation into the heart of central government policy making.

A question of evaluation

There are many kinds of policy evaluation but a simple means of categorising them is in terms of the question being asked and the timing of the question (Table 3.1). There are two basic evaluation questions, one descriptive: "Does the policy work?", the other analytic: "Why?". However, to avoid the teleological implications of "why?" questions, it is preferable to reformulate the second question as a "how?" question: "How does the policy work or not work?" Evaluations that address the first question are variously termed "summative", "program(me)" or "impact" evaluations and are typically quantitative (Orr, 1998; Greenberg and Schroder, 1997; Worthen *et al*., 1996). Those that focus on the second question are frequently called "formative" evaluations and are often qualitative (Patton, 2002; Pawson and Tilley, 1997, Yanow, 1999). However, a plethora of different terms has been used to describe formative evaluation that reflect subtly, and sometimes radically, different ontological positions (see below).

Increasingly, evaluations of public policies are combining elements of summative and formative evaluation (Gibson and Duncan, 2002). The fact that this is a comparatively recent development, particularly in the US, is somewhat perplexing since one would have thought it natural to want to know why a policy worked or did not. Perhaps evaluators thought that it was enough to know that a policy worked because it could then be continued or implemented elsewhere. (The term "demonstration project", commonly used in the US to describe programme evaluations suggests that policymakers have such great faith in the policy package being evaluated that the prospect of failure, and the need to analyse why, is seldom countenanced.) Alternatively, it may be because the methodologies designed to answer the two questions were developed in different parts of social science and have only come together belatedly through necessity or in recognition of the value of inter-disciplinarity in applied policy research (Ritchie and Lewis, 2003).

Table 3.1. **Questions of evaluation**

Time perspective	Evaluation question	Illustrative evaluation method(s)	Counterpart formative evaluation question	Illustrative formative evaluation approaches	Examples
Extensive past	What worked?	Meta-analysis Systematic review	How did it work?	Systematic review	Ashworth *et al.* (2002) (Case-study 4)
Past	Did the policy work?	Retrospective evaluation	How did it work/not work?	Retrospective interviews Participative judgement (Connoisseurship studies) Retrospective case-study	Huby and Dix (1992) (Case-study 3)
Present	Is this policy working?	Monitoring • Interrupted time series • Natural experiments	How is it working/not working?	Process studies Implementation evaluation Ethnography	
Present to future	Is there a problem?	Basic research Policy analysis	What is the problem?	Basic research Rapid reconnaissance	
Close future	Can we make this policy work?	Prototypes Micro-simulation	How can we make this policy work?	Theory of change Participative research Action research	Loumidis *et al.* (2001) (Case-study 2)
Future	Will this policy work?	Programme evaluation (Impact or summative evaluation) • Random assignment • Matched designs • Cohort designs • Statistical controls	How will it work/not work?	Theory of change Laboratory evaluation	Michalopoulos *et al.*, (2002) (Case-study 1) Hills *et al.* (2001) (Case-study 8)
Expansive future	What policy would work?	Prospective evaluation • Micro-simulation • Laboratory experimentation • Gaming	How would it work?	Laboratory evaluation Delphi consultation Gaming	Brewer *et al.*, 2001 (Case-study 5) Voyer *et al.*, 2002 (Case-study 6) Walker *et al.*, (1987) (Case-study 7)

Source: OECD.

The other dimension of the categorisation of evaluation studies relates to when in the policy cycle the two evaluation questions are asked. The tense used in the evaluation questions (present, past or future) will generally indicate whether the evaluation is conducted, concurrently, retrospectively or prospectively.

However, before discussing the implications of these distinctions, there is a prior question: "What is the problem or opportunity that requires an institutional policy response?"

Issues appear on the policy agenda for a range of reasons. These include: the occurrence of a crisis; the result of secular social, economic and/or political change; the successful activities of motivated individuals, interest groups, or editors; and errors and mistakes on the part of politicians or administrators (Hall *et al.,* 1975; Guess and Farnham, 2000). When an issue has emerged that may warrant a public policy response, the guidance offered in most policy handbooks is that basic research should be undertaken to delineate the nature of the problem and, where possible, to identify pathways of causality that may indicate points for policy intervention (Cm., 1999). However, handbooks are often ignored and this stage in the policy process is frequently omitted or undertaken only cursorily. When an issue is thought to be important and urgent, policy is often devised and implemented before the issue is well understood. Even where this is not the case, the research undertaken would normally be construed as applied research or policy analysis rather than evaluation. However, to the extent that much policy is iatrogenic, a response to the failure of pre-existing policy, this preliminary scoping research is likely to have an evaluative component: "If existing policy did not work, what were the reasons?" Moreover, if evaluation ever becomes a central element in the policy process, evaluative evidence will, in turn, be fundamental to any prospective policy review.

Programme evaluation

It is appropriate to begin discussion of the various kinds of evaluation with the model in which a policy is tested before full implementation since this is sometimes presented as the ideal-type evaluative strategy (Orr, 1998). In this case, the evaluation question is properly expressed in the simple future: "Will this policy work?" although the present tense ("Does this policy work?") is sometimes used, which has the unwarranted effect of turning a specific question into a general one that seems unbounded by time and place.

Addressing the question "Will it work?" usually involves conducting a programme evaluation or policy experiment (Greenberg and Schroder, 1997). Conceptually this is the most straightforward form of evaluation. Certain people are subjected to a policy intervention while others are not and the

outcome observed for both groups of people. Any difference in outcomes established between the two groups is interpreted as a measure of the impact of the policy, assuming all other things are held constant. Programme evaluation was pioneered in the USA with the celebrated negative income tax experiments of the early 1970s and remains the mainstay of the US policy evaluation industry.

Even so, there are complex issues of definition and implementation. First, the objectives of the policy need to be precisely defined and prioritised (this may itself be a spur to improved policy making which often makes do with political aspirations in lieu of objectives). The prioritisation of objectives is necessary because evaluation designs can rarely measure performance against multiple policy objectives with equal precision and are therefore usually devised to be most precise with respect to the most important objective.

Secondly, there should be agreement as to the degree of change that would constitute success. An aspiration to reduce the poverty rate, for example, has to be accompanied by a statement of the number of percentage points by which poverty is to be reduced. This is required so that samples used in the evaluation can be large enough to determine with adequate precision whether or not the policy has been a success.

Thirdly, some model or theory of change ought to be defined which would lead one to expect that the policy being implemented would indeed bring about the anticipated change. Such a model would allow appropriate outcome measures to be devised. Also, the better specified the model, the greater the number of intermediate outcomes that could be incorporated into the evaluation to test the model of change, and to provide diagnostic indicators when outcomes do not match up to expectations.

Finally, it is necessary to define a counterfactual, the situation that would have obtained had the policy to be evaluated not been introduced. The counterfactual provides a baseline against which the performance of the policy is to be assessed. It is usually inadequate simply to compare the pattern of outcomes before and after a policy is introduced since other features of the policy environment may change that influence the effectiveness of the policy. For example, a booming economy is likely to reduce poverty even in the absence of anti-poverty policies; in such circumstances, "other things would not remain constant" as required by the evaluative model. The role of the counterfactual is to partial out the effects of these other changes to isolate the impact of the policy alone.

The method that separates out the impact of a policy with minimum bias and maximum precision entails randomly assigning members of the policy target group to one of two subgroups: the so-called "action group" which is given access to the new policy and the "control group" which is not. Any

difference observed in the outcomes for the two groups can confidently be attributed to the effects of the policy since all other changes will, by definition because of the randomisation process, randomly influence both groups.

Although often treated as the gold standard in evaluation, randomisation is not without its limitations (Bottomley and Walker, 1996; Stafford *et al.*, 2002). One is the difficulty of securing political agreement. Politicians often object to random assignment for two related reasons. The first is that experimentation denies some people access to what may be considered a self-evident good. In reality, of course, the reason for evaluating a policy is uncertainty as to the benefits or, at least, the cost effectiveness of the policy. The second reason is that the presumed good is allocated at random rather than with respect to need or to the likelihood that people will gain from it. Given that it may be impossible to say in advance who will benefit most, if at all, from the policy intervention, random allocation is arguably as good a method as any of assigning scarce resources. Nevertheless, these concerns have proved to be a major obstacle to the use of randomised assignment in Britain.

There also technical limitations to random assignment. The most important arise when the policy to be evaluated is either intended to affect the system as a whole as well as individuals with in it, or when unintended consequences of the policy are likely operate at this level (Bottomley and Walker 1996). Take, for example, a policy to give welfare recipients a voucher that allows employers to offset some of the costs of employing them. In an experiment involving random assignment, only a proportion of welfare recipients would receive vouchers and they would enjoy a competitive advantage over other jobseekers that would disappear on full implementation when all jobseekers would be given a voucher. The effect of this, so called, queuing bias is to exaggerate the apparent effectiveness of a policy initiative. However, the system wide consequences, for example, a reduction in wage rates, are likely to be understated since not every welfare recipient receives a voucher; the partial equilibrium effect.

Unfortunately, there is no practical way of determining the scale of queuing bias or the partial equilibrium effect (Burtless and Orr, 1986). Quasi-experimental designs, in which a policy may be fully implemented in one jurisdiction and not in another, may allow for estimation of system wide effects, while also avoiding the need to allocate access to policies on a random basis. However, because no two areas are identical, the control of factors exogenous to the policy itself is much weaker.

Further difficulties associated with random assignment include ensuring that policy staff really do allocate access to clients at random and preventing members of the control group from surreptitiously obtaining the same services as the action group.

Case study 1

Programme evaluation: The Self Sufficiency Project – Michalopoulos *et al.*, 2002

A decade ago the Canadian Department of Employment and Immigration determined to investigate the effect that a "make work pay" strategy would have on the ability of long-term welfare recipients to make the transition into full-time employment. The Department therefore commissioned a 10-year demonstration project of a specially designed policy initiative based on a generous temporary earnings supplement, the Self-Sufficiency Project (SSP). This involved 9 000 lone-parent families in two provinces, New Brunswick and British Columbia.

To measure the impact of the SSP, a social experiment was conducted involving random assignment. A sample was drawn of welfare recipients who had been in receipt for more than a year, and one half randomly assigned to a programme group and offered the SSP supplement, while the remainder constituted the control group. Members of the programme group could receive the supplement for a maximum of three years.

Approximately 6 500 sampled welfare recipients were visited at home during which a 30-minute "baseline" survey was conducted. Respondents were the told that they had been selected to join the study and invited to sign a consent form after being told about the study and the principle of random assignment. The response rate for the baseline survey was around 90 per cent. Respondents were interviewed again 36 months and 54 months after random assignment, and administrative records from various government departments used to track their progress. The odds of being assigned to the programme group were 50:50 in both provinces with the exception of a 12 month period in New Brunswick when people were assigned equally to one of three groups: a programme group, a control group and a SSP-plus group in which participants were additionally offered job-search assistance and counselling.

Three subsidiary studies were nested in the design: the SSP applicant study, the SSP-plus study and the "Cliff" study. The first adopted an experimental design and entailed randomly assigning about 3 000 new applicants for welfare to a programme group, allowing them to receive the supplement 12 months after application, or a to control. The SSP-Plus study followed the experience of those offered SSP-plus, allowing for comparison with both the controls and those receiving the basic SSP. The cliff study examined the consequences of the withdrawal of the supplement after three years of receipt. Using administrative records, it followed the trajectories of 378 people identified in the 54-month follow-up survey to be approaching the end of their entitlement period. A sub-sample of 52 participants in this group were recruited to take part in a qualitative study comprising an initial focus group, followed by three telephone interviews, one before expiry of eligibility and two, respectively four months and eight months afterwards.

Case study 1 *(cont.)*

The evaluation demonstrated that SSP increased employment, earnings and income and reduced welfare use and poverty: programme group members received an average of $6 300 more in total income, including welfare payments, over the 54 month follow-up period. Combining the supplement with services (SSP-plus) helped people to find more stable employment than their counterparts in the control group. The social benefits of SSP outweighed the cost to government.

When randomisation proves impossible, the counterfactual has to be defined in other ways and numerous creative designs have evolved (Orr, 1998). These include matched group designs where members of the control group are chosen to be as similar as possible to members of the action (Loumidis *et al.*, 2001; Brice, 1996), cohort designs that seek to exploit situations in which successive generations follow the same trajectory (Smith *et al.*, 2001), interrupted time series designs that can be used when statistical series exist prior to policy implementation and various forms of natural experiment (Curington, 1994). In addition, analytic strategies have been designed, perhaps the most promising of which is propensity score matching, which attempt to define a counterfactual from variation within a non-randomised design. Indeed, Cullen and Hills (1996, p. 14) argue that quasi-experimental designs allied with statistical modelling "are the only practical solution to the unrealisable dream of total randomisation".

Prototypes

The policy prototype addresses a very different question from programme evaluation. With the decision to proceed to full implementation already taken in principle, the question asked is "How can we make this policy work?" The task in the prototype is therefore to fine-tune policy content to best effect and to determine the optimal mode of implementation. These aims place less emphasis on measuring outcome and more on understanding the process of implementation with the result that methodology is both eclectic and varied, including work-task analyses of the kind undertaken by operational researchers, and large scale, multi-method evaluations with both quasi-experimental, summative and formative components. Reliance on administrative data is also often heavy. Many of the major evaluations, the so-called "pilots", commissioned by the British Labour government since 1997, are more accurately called prototypes rather than programme evaluations.

The design of prototypes is often shaped by the rapid speed with which they are implemented and expected to report and by their closeness to the

Case study 2

Prototype: New Deal for Disabled People – Loumidis *et al.*, 2001

The New Deal for Disabled People personal adviser pilots is quite typical of the design of the policy evaluations commissioned under the UK Labour government after 1997. The aim was to determine whether recipients of Incapacity Benefit and certain other disability benefits would respond to an invitation to a work-focused interview with a personal adviser, and whether, with the assistance of personal advisers, more disabled people could secure and retain paid employment.

Two groups of six pilots were established in a range of different labour markets, the first group run by the Employment Service and the second by partnerships of public, private and voluntary organisations. The first set of pilots was initiated before commissioning of the evaluation. The design did not allow for randomised assignment since the Personal Adviser Service was to be made available to all eligible disabled people in the pilot areas. Therefore, the invitation to tender suggested the establishment of 12 comparison areas. In fact, the Department also had aspirations to generate base-line statistics against which to assess the impact of any national implementation. It therefore commissioned a design that included, in addition to interviews with applicants to the New Deal programme, interviews with a national sample of disabled people drawn from outside the pilot areas but stratified according to the same criteria as were used to select the pilot areas. This national survey was intended to be used to establish the counterfactual against which the effectiveness of the Personal Advisers is to be assessed. A comprehensive programme of process evaluation accompanied the impact analysis.

A critical issue in the design of all policy evaluations is the anticipated size of any effect. When large, sample sizes can be comparatively small. However, should the actual effect turn out to be much smaller than expected, the power and precision of a design can be severely tested. In the case of the New Deal for Disabled People pilots, resource constraints served to limit attainable sample sizes (approximately 3 000 in total) while the take-up of the scheme proved to be much lower than expected.

While the New Deal for Disabled People pilots were commissioned explicitly to inform decisions about the possibility subsequent national implementation, it was always intended that such decisions should be taken half way through the two year pilot and before the results of impact analyses were available. In such circumstances, the advance of policy did not fully benefit from sophisticated evaluation, or, at least, not in the short-term.

policy process. Interim and repeated reporting is normal and often linked to a staged rollout of policy, with evaluation results being used to alter implementation in a way reminiscent of action research. Indeed, in Britain full policy implementation has often preceded the results of the prototype evaluation becoming available (Walker, 2001).

Some prototypes in Britain have commenced with very limited prescription of policy content. Agencies on the ground were charged to develop this within a resource framework and descriptive accounts of the policies evolved reported to policymakers with or without a research-based commentary (Walker, 2000a).

Monitoring and retrospective evaluation

Changing the evaluation question from the future tense to the present, "Is policy working?" or the past, "Did policy work?", means that defining a secure counterfactual is all but impossible for the simple reason that the policy is already available to everybody as the result of full implementation.[1] In these scenarios, evaluators turn to monitoring and retrospective evaluation.

Monitoring and retrospective evaluation remain popular approaches despite their obvious imitations. This is partly because they form part of the normal process of policy audit, which seeks to establish who has receives the service and at what cost, often by reference to administrative information. Retrospective evaluation may also be triggered by suspicion, often aroused by monitoring, that the policy is not working well. These modes of evaluation do not require the same level of institutional commitment to evidence based policy making as programme evaluation. They are not, for example, located on the critical path from policy idea to policy introduction that demands policymakers rein back their enthusiasm for implementation to await the outcome of a lengthy evaluation. Monitoring and retrospective evaluation are also generally cheaper than programme evaluation, but tend to answer different questions in different ways.

The lack of a secure counterfactual typically means that greater emphasis is given to resource inputs and their conversion into service provision (administrative efficiency) than to the contribution of service delivery to meeting policy objectives (effectiveness). This is the province of operational researchers, official statisticians and sociologists with an interest in institutions and policy implementation rather than economists who, especially in the US, have led the development of programme evaluations. Frequently, the results of monitoring find their way into the public domain as compendia of discrete statistics (Cm. 2002) rather than as rounded assessments of particular policies, which is principally the preserve of retrospective evaluations.

The methodologies employed in retrospective evaluations tend to be eclectic and adverse circumstances can stimulate creative designs. Pluralistic approaches are often used in which the experiences and opinions of key actors in the policy implementation are collated and triangulated to reach an overall judgement on policy effectiveness. Personal interview surveys may be used to solicit the views of policy recipients, qualitative interviews conducted with administrators and other interest groups and observational techniques used at the point of service delivery. These accounts can provide irrefutable evidence about the efficiency or lack of efficiency of implementation and provide a sound basis for reform.

A particular focus is often on targeting, since a *prima facie* case can usually be made that if a policy is not reaching its target population it is unlikely to be particularly effective. Two aspects are important: first, the proportion of policy recipients who receive services unnecessarily because of poor policy design, mal-administration or fraud, and secondly, take up: the proportion of the eligible population that actually receives the service (Knapp, 1984; Walker, 2004). The first can usually be informed by assembling administrative statistics especially if judgements about who receives services "unnecessarily" are made in relation to the programme specification rather than to policy outcomes (which would might require the definition of a counterfactual). Specification of take-up often poses greater difficulty since eligible non-claimants are usually invisible to the administration and hard to track down empirically.

Where retrospective evaluations have sought to assess policy impact they have typically used one or more of three approaches: trend analysis, quasi-experimentation and reportage. At it simplest, the first approach entails searching for an inflection in a time-series variable that coincides with its introduction of the policy. If there is confidence that the variable is likely to be affected by the policy, and an inflection is apparent and in the right direction, the policy is presumed to have had an effect, the abruptness of the inflection indicates the size of the effect. More sophisticated analyses use time-series regression or other simulation techniques to define a counterfactual by predicting the trend of the variable in the absence of the policy and comparing the prediction with the actual trend (White and Riley, 2002). The success of this approach depends on the reliability of the trend variable, the precision of the regression predictions and the stability of the relationships before and after implementation of the policy.

The second approach depends on the identifying an ostensibly similar group who are not affected by the policy to serve as a counterfactual and comparing the experience of this group with that of people targeted by the policy. Hasluck *et al.* (2000), for example, used mothers in couples as a counterfactual for lone parents targeted in a welfare to work policy. The

potential of this approach is limited by the difficulty of finding a counterfactual and the degree of congruence between the counterfactual and the policy target group.

The third approach generally dispenses with an explicit counterfactual and relies on policy actors' assessment of effectiveness (Thornton and Corden, 2002). They may be asked, using either open or structured questions, how effective they consider the policy to be overall, and perhaps how effective they judge its component parts to be. Absolute judgements of this kind are sometimes complemented or replaced by relative measures that might encourage respondents to compare the performance of the policy with that which it replaced which is akin to using the former policy as a perceptual baseline or counterfactual. The value of such assessments depends crucially of the knowledge and critical judgement of respondents but where there is a large measure of consensus among respondents, they may have considerable reliability (Walker with Williams, 1987). Nevertheless, because of their subjective nature, they are usually one element in a multi-method retrospective evaluation.

It is important to note that while monitoring and retrospective evaluations are necessarily weak on establishing cost effectiveness, programme evaluations have traditionally paid little attention to administrative efficiency. This is not a necessary consequence of the methodology but the result of research priorities. Programme evaluations have often successfully demonstrated whether a policy works but less frequently how or why.

It is appropriate to consider prospective and meta-evaluation together even though prospective evaluation looks to the future and meta-analysis draws insight from the past. They lie adjacent at the point in the policy cycle when the challenge is to consider what policy might be implemented in response to a defined problem – prospective evaluation – and the first step is to consider what worked in similar settings in the past – meta-evaluation.

Meta-evaluation

Meta-analysis is a set of statistical techniques that have been developed to aggregate and summarise results from existing studies (Lipsey and Wilson, 2001). Developed in psychology and widely used in medical research, meta-analysis is often coupled with systematic review, a set of procedures used to identify as many relevant studies as possible (so as to minimise selection bias) and to evaluate them in terms of quality of design, execution and analysis (Glass, 1976). Formal meta-evaluation, meta-analysis applied to policy evaluations, usually entails a regression analysis in which the dependent variable is the programme outcomes from a number of policy experiments and the set of independent variables describe programme content and

Case study 3

Retrospective evalation: Social Fund evaluation – Huby and Dix, 1992

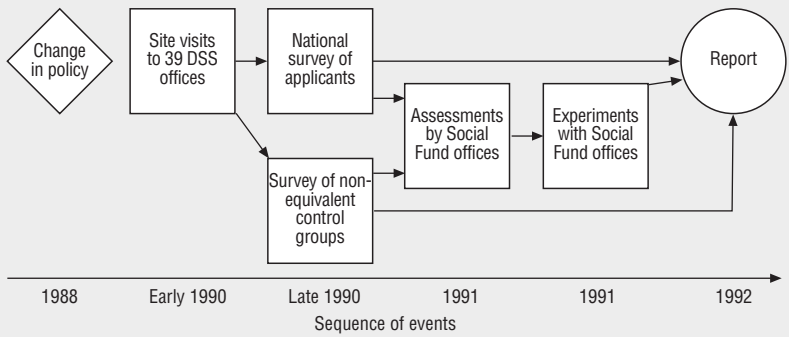
The Social Fund, introduced in the UK in 1997, is a system of largely discretionary loans and grants to meet the one-off needs of people living on social assistance which are paid from a fixed budget allocated to local offices.

The evaluation sought to determine the extent to which the Social Fund was targeted on people who were most in need, an objective which was complicated by the discretionary, budget driven nature of the scheme. Only social fund officers using discretion could determine whether a person was eligible, but even their discretion was fettered by the budgetary constraints.

The approach adopted after competitive tender was to define four non-equivalent comparison groups (Figure 3.1):

1. successful applicants of Social Fund;
2. applicants who had been refused an award;
3. income Support recipients who had not applied to the Social Fund to establish how far need was going unmet; and
4. recipients of Housing Benefit, who were not entitled to apply to the Social Fund, but who were included as proxies for low income families.

Figure 3.1. **Evaluating the Social Fund Prospective and meta-evaluation**



Source: OECD.

administration and the environmental context (Ashworth *et al.*, 2002b). The regression analysis seeks to establish which features of the policy and its implementation are statistically associated with observed variations in

programme outcomes, that is to determine which aspects of policy and implementation work best.

To undertake a meta-evaluation, information on programme outcomes, including effect size and associated standard errors, is extracted from reports of an exhaustive set of relevant policy evaluations along with details of programme design, implementation and the local policy environments. In the regressions, estimates of the programme effect size (that is, the dependent variable) are weighted by their standard errors to take into account the different precision with which the effect size estimates were obtained.

Meta-evaluation does more than summarise the existing evidence about what works, it offers insight into how policies work. Whereas a single programme (or impact) analysis will determine the cost effectiveness of a particular policy, and associated formative (process) analyses will point to aspects of its delivery that may contribute to its effectiveness, meta-analysis can provide a quantitative estimate of the added value of each feature of a policy and its administration, both in isolation and in combination. Unlike a single evaluation, meta-analysis can also establish whether a particular policy would work as well in a different location or, indeed, whether the apparent success of a policy implementation is due in large measure to the idiosyncratic nature of its policy environment.

The value of meta-evaluation will vary according to the number and quality of relevant evaluations available and the degree of correspondence between the current policy conundrum and those addressed by previous policies that have been evaluated. Currently the former limitation is very severe since the number of policy evaluations conducted in any particular policy domain and particularly within the same jurisdiction, is likely to be small. However, this problem should reduce with time as more policy implementations are systematically evaluated although there will always be issues about the comparability of evaluations and the variables used in them that can only be partly addressed by calls for best practice in the design and reporting of evaluations.

Prospective evaluation – micro-simulation

In principle, meta-evaluation can be used prospectively, rather than retrospectively, to tackle the question "What policy would work?" Once a regression model has been developed, the characteristics of various policy options and settings can be substituted into the equation to establish their likely outcome and relative cost effectiveness. However, data limitations mean that this form of policy simulation has yet to have the same impact as that based on static micro-simulation models and, more recently, dynamic micro-simulation. A further approach to prospective evaluation that is likely to play an increased role is laboratory experimentation.

Case study 4

Meta-evaluation: Discovering what works best in welfare provision – Ashworth *et al.*, 2002

This study assembled a database comprising material gleaned from all 24 mandatory US welfare to work evaluations implemented between 1982 and 1996 that had used random assignment. The database contains information covering impact estimates and programme characteristics for 64 policy programmes extracted from published reports, supplemented by data on labour market conditions compiled from other sources.

The objectives of the meta-analysis were to identify which programme characteristics were most important in explaining variation in the success of programmes in increasing earnings and reducing welfare receipt; to establish for how long the impacts lasted; and to ascertain whether programme outcomes were in any way affected by differences in labour market conditions, or the composition of the welfare caseload.

The analysis employed a weighted least-squares regression approach with the programme impact of interest entered as the dependent variable. The weighting adjusted the programme impact variable for the sampling variance of each of the original evaluations. The weight was calculated as the standard error of the impact estimates so that impacts with smaller variances were given a larger weight to reflect their robustness.

The analysis confirmed the widely accepted view that interventions that emphasised measures to encourage people to return to work quickly, rather than to engage in human capital acquisition, reduced welfare rolls most and resulted in higher earnings gains. It also indicated that statistical differences between the original action groups and controls attributable to the impact of policies lasted for several years. However, the performance of policy interventions was sensitive to differences in the local environment and to variations in the characteristics of the welfare caseload and these factors helped to explain the apparent outstanding success of certain programmes.

Micro-simulation models attempt to predict the likely impact of a policy change on individual persons or families, aggregating the individual effects to provide estimates of the total impact (Klevmarken, 1997; Plum, 1998). Some of the simplest and most successful models have been used to assess the impact of tax and benefit changes. Before the advent of modelling, the impact of such changes would have been discussed in terms of the impact on an "average" family. Since most families do not closely resemble the average, such an approach was often misleading. Modelling the impact on everyone in a representative sample of families on the basis of their income and family

circumstances, as assessed in survey interviews (or, where possible, using details from administrative records), has proved to be much more successful. In essence, the survey data serves as a model of the target population comprising individuals each with sets of endogenous and exogenous attributes, the former postulated to be susceptible to change due to the new policy. This initial model serves as the counterfactual, the product of existing policies.

The endogenous attributes are then "updated" according to a set of transition probabilities that simulate the effect of the policy change, which are then "grossed up" according to sampling fractions to provide population estimates. Early models were static in that they provided only the immediate, first round effects of the policy change but by incorporating behavioural assumptions into micro-simulation models, it is possible to take account of second and subsequent round effects. So, for example, tax and benefit models will allow for changes in expenditure patterns resulting from alterations to indirect taxes, and changes in employment brought about by modifications in the incentive structure caused by reform of benefit levels or tax rates.

The success of micro-simulation depends on the quality of data and modelling. The modelling is in turn dependent on prior understanding of the system being modelled and the appropriateness and specificity of underlying theory. Of critical important is the validity of the behavioural assumptions embedded in the modelling; these are typically derived from a judicious mix of often quite ill-developed theory and empirical evidence. Moreover, the more major the change of policy, the more likely it is that behavioural assumptions will no longer be relevant, especially if the model has been calibrated with reference to behavioural changes observed in a radically different policy environment. Experience suggests that modelling can benefit from open discussion of the methods and assumptions and even from competition from different teams of modellers (Cabinet Office, 2000).

Developing a model of a complex policy environment is a lengthy procedure with high associated costs but these may be offset against the longevity of the resultant models.

Prospective evaluation – experimentation

Unlike policy simulation, laboratory experimentation, which offers the second, probably generally complementary, route to prospective evaluation,[2] can be set up rapidly at comparatively limited cost. Laboratory experimentation takes one of two basic forms but with great scope for mixing the two (Huby and Dix, 1992). The first, *behavioural experimentation*, adopts a positivist approach seeking to measure individual behaviour in a controlled setting. The second, here termed *laboratory evaluation*, creates an artificial setting in which respondents are stimulated to reflect on policy options.

Case-study 5

Prospective micro-simulation: An assessment of the new Tax Credits – Brewer *et al.*, 2001

The UK government is to implement major reforms to the tax and benefit system in 2003 with the introduction of Child Tax Credit, which merges much financial support available to children into a single income related scheme, and Working Tax Credit which will provide financial assistance to employed adults in low paid work.

Unlike the other case-studies, this prospective evaluation was not directly funded by government although technical assistance was provided to the researchers. The objective of the work was to establish, based on details that had been released about the proposed reforms:

- the number of likely beneficiaries;
- the impact on poverty rates for different types of family; and
- the effect on work incentives brought about by changes in the budget constraint (the result of the trade-off between wages received through engaging in extra work and the consequent loss in financial assistance as a result of increased income).

The analysis was undertaken using, TAXBEN, a dynamic micro-simulation model of the UK tax and benefit system developed by the Institute for Fiscal Studies. Rather than simply calculate the immediate first-order financial effects of proposed tax reforms, TAXBEN integrates essentially static models of the tax-benefit system with behavioural models to enhance understanding of the "second-round effects" of policy changes. TAXBEN is therefore integrated with a labour supply model to provide estimates of the effects of changes in benefits or direct taxation on incentives to work. This involves calculating individual budget constraints for each person in the sample and simulating the desired level of hours worked under different tax regimes.

The evaluation suggested that 5.7 million families would be entitled to the new Child Tax Credit, that poverty rates among families with children would be reduced by three percentage points, and that 1.1 million families with moderately higher initial income would be made worse off due the reforms. Analysis of the budget constraints implied by the reform showed, for example, that the first person in a couple without children would be better off not working at all than working for between 16 and 20 hours a week in a job paying the national minimum wage.

Behavioural experimentation is typically used to establish a policy principle rather than to evaluate a tightly defined policy option. A key concern in devising policy is the likely reaction of the target population; this is the

behavioural response that some authors have identified as the Achilles heal of micro-simulation. Experimentation seeks to establish this response directly rather than inferring it from the observed outcomes of past behaviour or relying on self-reported accounts elicited from survey interviews. Respondents are asked to make real decisions that reveal the preferences that policymakers are interested in.

The techniques of behavioural experimentation are derived from experimental economics but with certain crucial differences. In experimental economics, respondents (experimental subjects) are presented with choices in which it is costly for them to misrepresent a preference because they stand to forego real money if they do so (Blondel *et al.,* 2000; Eckel and Grossman, 2001). Using so called compensated questions or options, respondents may, for example, be asked to choose between having £10 to spend immediately or £15 in six months time. Because circumstances are controlled, any variation in the preferences expressed that might result from this cause is eliminated.[3] When behavioural experimentation is applied to policy evaluation, circumstances are again tightly controlled but the small financial stakes are replaced by large ones that may approximate to the cost of particular policy options. Respondents choose between sets of options designed to establish the structure of their preferences simulating the policy design knowing that one their choices, selected randomly by researchers, will be honoured by a cash payment or voucher equal to the relevant stake. Moreover, instead of typically using students and seeking to define average relationships, respondents are drawn from the policy target population with attention often focussed on variation in the pattern of revealed preferences.

The art of behavioural experimentation is to invent compensated options that inform key decisions about the structure of a policy designed to influence behaviour. The closer the correspondence between the compensated options and features of policy intended to trigger behavioural change, the better experimental results are likely to predict policy outcome, especially where respondents are representative of the target population.

Laboratory evaluation is a specialist form of policy consultation that engages various groups of policy actors in critically evaluating policy options. It may be thought of as a formative analogue of behavioural experimentation, focussing on the critical analysis rather than empirical testing of the logic underlying the policy design. A common format is the extended creativity group in which sampled or otherwise selected policy actors are brought together for a day or two, informed of the policy objectives and options and their opinions sought after and exchanged in moderated groups of various size and composition or, as appropriate, elicited individually (Walker, 1985b). The aim may variously be to exploit the expertise of the group to conduct a SWOT (strengths, weaknesses, opportunities, threats) analysis, identifying

Case-study 6

Prospective laboratory experimentation: Fostering adult education – Voyer *et al.*, 2002

The Canadian government wishes to ensure that at least a million additional adults pursue learning opportunities over the next five years and the Human Resources Department is currently examining various options to meet this objective. The Department needed, in particular, to establish by how much various types of programme or financial incentives would affect the behaviour of the adult population. To address this problem it commissioned a large-scale laboratory experiment.

The experiment was designed first, to determine what types of government assistance best serve the policy objective of increasing human capital investment among adults from different socio-economic backgrounds and second, to establish the barriers that prevented adults from engaging in education .

The project comprised a representative interview survey of adults to ascertain demographic, socio-economic, behavioural and attitudinal characteristics; a series of individual choice-questions involving monetary compensation to capture revealed preferences experimentally; and a literacy assessment to measure ability and perceived ability to learn.

Table 3.2. **Preference for student loans**

	Choice A Cash	Choice B Up to this amount in student loan for full-time or part-time education
Decision 21	$100 or	$100
Decision 22	$100	$200
Decision 23	$100	$300
Decision 24	$100	$400
Decision 25	$100	$500
Decision 26	$100	$600
Decision 27	$100	$700
Decision 28	$100	$800
Decision 29	$100	$900
Decision 30	$100	$1 000

The choice questions were designed so that the cash alternative to each education choice under the various types of financial assistance remained the same. As a result, it was possible to observe the level of financial assistance at which individuals were willing to switch from one type of assistance to another. Linking the survey findings to the experimental data enabled calculation of the proportion of people who would invest in education given the various forms of assistance, while regression analysis indicated how choices varied by socio-economic status, fear of failure, loan aversion and other personal characteristics.

Source: OECD.

Case-study 6 *(cont.)*

The choice-questions involved an offer to enrol on an educational course with support that varied in kind and level with the intention of eliciting preferences for education when financed by a grant, a loan, an income sensitive loan or by subsidised savings. The choice-questions for loans are given in Table 3.1. During the experiment respondents chose between A and B for each question and, on completion, one question was selected at random and each respondent provided with the pay-off corresponding to the choice made in the selected question. For instance, an individual who selected B for decision 28 would receive a loan of $1 600 to enrol in education.

potential strengths and weaknesses of the policy together with opportunities and threats to it, to seek a consensual response or to define the degree of support or opposition. An alternative goal may be to develop or refine the policy design directly in the research setting, perhaps iteratively with groups meeting repeatedly to work up particular aspects or with modified designs being passed from group to group for analytic evaluation. The Delphi technique, developed by the Rand corporation in the late 1970s, is a variant in which policy analysis and development take place through an iterative process of consultation (Goodyear-Smith and Farnell, 2001).

The value of laboratory evaluation is dependent on the expertise of respondents and the skill of the research moderators to capitalise on it. It may be most useful when the policy changes proposed are path dependent and current experience is likely to be germane and least so when a change in policy paradigm is proposed. Even so, the technique may still have worth in the latter setting, informing the management of change.

Formative evaluation

There is inadequate space to do justice to formative evaluation which has a rich history and is characterised by diversity of approach born from different ontological positions and creative responses to particular policy implementations (see Ritchie and Lewis, 2003; Patton, 2002 for reviews). For each summative evaluation question, there is a set of corresponding formative ones that seek explanations for, or understanding of, the outcomes of policy (Table 3.1). These questions lend themselves to qualitative research and so formative evaluation is characterised by reliance, albeit not exclusive reliance, on qualitative techniques such as depth interviews, focus groups, observation and case study.

Case-study 7

Prospective laboratory evaluation: Discussion about Housing Benefit reform – Walker *et al.*, 1987

The implementation of the 1982/3 Housing Benefit scheme by local authorities in the UK proved to be exceedingly problematic, which led the Thatcher government to commission an independent review of Housing Benefit. This, in turn, informed Green and White Papers on social security reform (Cmnd, 1985a; b).

Research for the Review Team had demonstrated that some of the administrative difficulties were inherent in the structure of the scheme, which was the result of the merger of earlier schemes within a common framework that had largely failed to unify them (Walker, 1985b). This led to great complexity and made effective functioning of the scheme dependent in a degree of inter-agency liaison that was hard to sustain. The White Paper addressed many of these issues but left key aspects of the reform to be decided. The Department chose to use research to tap the expertise of those directly responsible for administering Housing Benefit in local authorities so as to ensure that new procedures would work.

The research was let by negotiated single tender. Senior housing benefit officials from a stratified random sample of 66 local authorities participated in a series of day-long research workshops, during which, at various times, participants worked in groups varying in size from two to eight people depending on the objective. Two groups of officials met on two occasions a week apart. Their task was to review the government's detailed proposals and, on the second occasion, to suggest alternatives. Six other groups met once to evaluate the full set of proposals. Two self-completion questionnaires were administered, one before and one after the workshops. The first gathered information on performance indicators, the second mainly elicited a reflective response to all the policy proposals as refined during the course of the workshops. Follow-up telephone interviews were also conducted with a view to enriching the written responses.

One aspect of the consultation concerned the change in the definition of income used to assess a person's entitlement from gross to net. Initially the government proposed that, rather than attempting to collect information on actual net income, a formula be used to assess a proxy “notional” net income. Participants rejected the government’s proposal as being too complex for rough and ready assessment but not sensitive enough to ensure equity. Over the course of the research, they moved increasing in favour of assessment of actual net income, the approach that the government finally adopted.

The tense in which each question is asked again influences the evaluative design but not to the same extent as with summative evaluation. Where the evaluation is retrospective, formative evaluators will necessarily rely more on documentary and secondary evidence than is the case with evaluation that is ongoing where both observational and action research are possible, the latter involving the engagement of researchers with policy actors in real-time to assist them in refining policy design and implementation. Similarly, prospective evaluation is more likely to rely on projective techniques, simulation and role-play, perhaps used together as in the laboratory evaluation discussed above (Walker *et al.,* 1987; Case-study 7). However, the broad approach of most formative evaluation is pluralistic and investigative, assembling whatever relevant information it is possible to obtain, and comparing and contrasting insights gleaned from the perspectives of different policy actors and sources of data.

It is important to hold tight to this simple formulation because qualitative analysis is grounded in numerous, often competing, traditions. Patton (2002, p. 132-3) lists 16 established traditions including ethnography (concerned to describe cultures), constructivism (which seeks to explain how people construct reality), symbolic interactionalism (that investigates the meaning individually and collectively given to people's interactions) and narrative analysis (that looks to people's accounts of their life experiences better to understand their beliefs and actions.) Each tradition differs in the response given to the following profound questions:

- What do we believe about the nature of reality?
- How do we know what we know?
- How should we study the world?
- What is worth knowing?
- What questions should we ask?
- How should we personally engage in the enquiry?

To this list, Patton adds "pragmatism" in which choice of method is separated from epistemology and matched to specific research questions. Walker (2003) argues that the ontological position of evaluators who take this stance:

> "is likely to correspond to 'subtle realism' (Hammersley, 1992), accepting that a diverse and multi-faceted world exists independently of subjective understanding but believing it to be accessible via respondents' interpretations. They will probably strive to be neutral and objective at each stage in the research process and thereby to generate findings that are valid and reliable, ones that are true to the beliefs and understandings of their respondents and potentially generalisable to other settings. They will seek detailed descriptions of the realities as understood by their

respondents, but clearly delineate between these interpretations and their own in analysis, the process of making sense of complexity by simplification and structuring."

Pragmatists are likely to accept a role for both quantitative and qualitative methodologies and to appreciate the complementary nature of summative and formative evaluation, the latter enriching understanding of the reasons that policies work or do not work well (Ritchie and Lewis, 2003; Walker, 1985a). However, evaluators working in some other traditions consider summative evaluation, especially that based on experimentation, to be inherently flawed and inferior to formative evaluation. Pawson and Tilley (1997), advocates of "realistic evaluation", which is grounded in scientific realist philosophy (Lakatos, 1970; Harré, 1986), adopt this stance.

Even within the pragmatist framework, there are myriads of evaluative models and approaches, the key ones being listed in Table 3.3, adapted from Patton (2002) who provides a brief account of each one. The models reflect the different purposes of evaluation within which different approaches may be used. In the classic objective-orientated model, a policy is evaluated with respect to the objectives set by the policy architects. In such a case, depending on the approach chosen, the formative evaluation seeks to investigate how the policy is implemented, to understand better how interactions between the various policy actors might affect outcomes, or to triangulate different perspectives on the working and effectiveness of the policy. Goal-free evaluation does not prioritise those outcomes that are directly linked to the policy objectives but explores a wide range of intended and unintended outcomes and the antecedent processes. Transaction evaluation emphasises the different perspectives of all the policy actors, policy makers, administrators, field level staff, clients, etc., and may entail direct interaction between evaluators and some or all of these groups with a view to working collaboratively to improve implementation and outcomes. Utilisation-focused evaluation prioritises understanding of the policy implementation process in the belief that the way that policy is shaped by delivery critically influences its likely success. Finally, connoisseurship studies place the evaluator(s) in the role of judge, sifting and evaluating evidence in the light of professional and personal expertise.

Turning to the specific approaches listed in Table 3.3, four warrant special mention. First, process studies epitomise formative evaluation and frequently accompany summative evaluations. They seek to elucidate the internal dynamics of how policies operate. The questions addressed by process studies include the following:

- How is the policy delivered in practice?
- How do clients enter the system and what happens to them as they progress through the system?

Table 3.3. **Formative evaluation: models and applications**

Evaluation models
Objective-orientated evaluation
Goal-free evaluation
Transaction models: Responsive and illuminative evaluation
Connoisseurship studies
Utilization-focused evaluation
Evaluation applications
Outcomes evaluation
Evaluating individualized outcomes
Process studies
Implementation evaluation
Theories of change/action; logic models
Evaluability assessments
Comparing programs: focus on diversity
Prevention evaluation
Documenting development over time and investigating system changes
Interactive and participatory applications
Personalizing and humanizing evaluation
Harmonizing program and evaluation values
Developmental applications: action research, action learning, reflective practice, and learning organisations
Appreciative inquiry
Participatory research and evaluation: valuing and facilitating collaboration
Supporting democratic dialogue and deliberation
Supporting democracy through process use: helping the citizenry weigh evidence and think evaluatively
Applications focused on quality
Understanding and illuminating quality
Quality assurance
Special applications
Unobtrusive measures
State-of-the-art considerations: lack of proven quantitative instrumentation
Confirmatory and elucidating research: adding depth, detail, and meaning to quantitative analyses
Rapid reconnaissance
Capturing and communicating stories
Legislative monitoring and auditing
Futuring applications: anticipatory research and prospective policy analysis
Breaking the routine: generating new insights

Source: Adapted from Patton (2002).

- What experiences does the policy generate for the various policy actors?
- How do staff, clients and resources interact?
- How do beliefs and actions link to outcomes?
- How do the actors account for the policy outcomes?
- What are the strengths and weakness of the policy that are identified by policy actors, and what is the nature and degree of consensus about how well the policy works?

Secondly, theory of change (Rogers *et al.*, 2000; Weiss, 1996; Chen, 1990) or theory of action (Patton 2002) approaches warrant attention because of the

profound influence that they are beginning to have on all kinds of evaluation, summative and formative. Their shared characteristic is the aim of seeking people's perception of the sequence of causation thought to link policy inputs, implementation and outcomes. They differ in terms of the people whose views are sought, in the timing of this quest and in the use made of the results. Some will elicit a causal model from the policy designers early in the evaluative process, and design a methodology to test or establish the validity of the theory in practice. Others will seek causal models from different groups of policy actors once the policy has been implemented in the belief that lack of congruity in perceptions may help to account for policy failings. Yet, others engage with policy actors to devise or refine causal models with a view to developing and improving policy design and implementation. The major contribution of the theory of change approaches is to draw attention to the need to make explicit and transparent the black box of policy. While it may be possible to establish whether a policy is a success without a clear understanding, or at least hypotheses, of *how* policy is to work, without such knowledge it is very difficult to fine-tune policy or to derive lessons that might have applicability in different arena.

Third and fourthly, formative evaluation is often used in situations when quantitative summative evaluation is impossible or inappropriate. Patton (2002) notes "state of the art" situations when no acceptable, valid or reliable measures exist. Similarly, qualitative research can be used to confirm, qualify, interpret, illuminate and illustrate summative evaluations; when comprehensive and/or not explicitly subservient to the summative evaluation, such research warrants the label "formative evaluation".

Returning to the simple typology introduced in Table 3.1, this provides a summary of the argument up to this point. It identifies the different purposes of policy evaluation, linking these to the kinds of knowledge required for policy making. It suggests that different evaluation designs are well suited to informing different policy questions; "suggests" because evaluation design is a profoundly creative process in which there is seldom a single ideal methodological solution. Finally, it points to the need for evaluative evidence at each stage in policymaking cycle. To achieve this requires a sustained political commitment to building policy based on evidence and an infrastructure able to deliver appropriate information at the most opportune time.

Finally, the evaluators felt able to comment on the effectiveness of the schemes. The targets of the number of placements that schemes forecast were abandoned, as most were judged unrealistic, and the evaluation focussed, instead, on the percentage of clients recruited who moved into work and how selective the schemes were in their recruitment. Of the schemes that were judged to be most "successful", five were selective on job readiness criteria

Case study 8

A formative prototype: new deal for disabled people – innovatory schemes – Hills *et al.*, 2001

The UK government in 1997 piloted initially two variants of New Deal for Disabled People. One (Case-study 2) employed personal advisors, the other sought to identify and test different approaches to helping people move into or remain in work. After a competitive tender, 24 separate schemes were established and evaluated using a largely qualitative methodology.

The evaluation of the innovatory schemes pilot was commissioned by the Department of Social Security in the basis of a competitive tender. A range of different methodological devices were employed including:

- an audit of each scheme involving site visits at the start and conclusion of the two year programme, documentary analysis and interviews with staff, participants and partner organisations;
- thematic case-studies around themes such as work with employers and relationships with partners;
- scheme based case-studies to explore relationships between the schemes and the wider organisation and the pathways followed by clients through the service;
- analysis of monitoring data provided to the Department of Social Security by the schemes; and
- dissemination workshops used to validate the learning by presenting evaluation findings to representatives of the schemes.

During the course of the evaluation, the evaluators developed a "pathway" model that helped to map the different kinds of activities in which the schemes were involved. This model, though not used as a tool by the schemes, indicated the need for both mobilising and supporting, and matching and mediating activities, to take place with both clients and employers.

The evaluation was able to describe and comment on the trajectories followed by clients, the procedures adopted for developing partnership working with other agencies, management styles and tasks and the financial viability of schemes. It also sought to identify the types of client targeted by schemes and to examine the nature of the relationships between scheme, clients and employers.

and four were not whereas schemes that were comparatively "unsuccessful" were not so selective in their recruitment procedures.

Threats to evaluation

The typology in Table 3.1 also presents an ideal-case scenario in which policy evaluation would be conducted and *used* to inform decisions at each point in the policymaking cycle. In reality, as many authors have noted, most policies are not formally evaluated and even when they are, the results may never be used or simply be exploited to justify a prior decision (Bulmer 1986; Walker, 2000; Weiss, 1992; Wilensky, 1997). Rather than simply add to this disappointing litany, Britain post 1997 is taken as a case study to illustrate some of the challenges faced by an administration that is committed to grounding policy making in evidence. Then, in the final section, this case study to used to reflect on ways of ensuring, or at of increasing the likelihood, that good evaluative evidence will be produced and effectively utilised.

As an important aside, it is worth recognising the methodological difficulties of establishing whether any particular evaluation has had an effect on policy (Weiss, 1998). Invariably there is no counterfactual demonstrating what would have happened in the absence of the evaluation so that one is generally forced to be reliant on reportage (Davies *et al.,* 2000; Greenberg *et al.,* 1999; Cabinet Office, 2000). The nature of government bureaucracy with specialisation of function means that few if any people are in a position to view the policy process as a whole and to be able to isolate the impact of evaluative research. Many of the key players will also have had reason either to overplay or down play the role of evidence. Moreover, if the task of attributing a change in policy to research evidence is difficult, the problem is multiplied when any evaluative evidence produced supports the political thrust of policy such that no change in policy is needed or evident. Furthermore, studies of the impact of research on policy (as opposed to the role of evaluative evidence *per se*), suggests the processes are often indirect and effective only in long term, seeping through the corridors of power as water permeates limestone, slowly changing collective understanding of the issues (Thomas, 1987; Weiss, 1980).

Policy environments

Turning to the United Kingdom as a case study, there has been a radical shift over the last twenty years towards prioritising evidence in policymaking (Davies *et al.,* 2000; Walker, 1997). This development was stimulated in the 1980s by the emergence of the so-called "new public management" (Hood, 1991; Stewart, 1996) and the associated introduction of managerial doctrines and techniques to improve the monitoring, control and evaluation of performance (Carter and Greer, 1993). With the principles established, in 1988 HM Treasury published the first edition of an enduring guide to policy evaluation for managers. This included an "evaluation framework" that led

the policy manager through a discussion of the purpose, nature and execution of evaluation (Treasury, 1988). For the most part, the guide was limited to the evaluation of existing policies, although reference was made to the role of evaluation in policy appraisal.

These changes were accompanied by a substantial increase in policy evaluation during the 1980s and 1990s (Walker, 2000b). Typically, this evaluation was retrospective rather than prospective. It therefore relied on pluralistic approaches rather than the random assignment methodology that was by then becoming the mainstay of policy evaluation in the US. The reasons for this different approach were mainly structural. Britain is highly centralised with many policies being implemented uniformly directly by agencies of central government, whereas the US federal system devolves much more policy making to the state level. Until 1996, varying social security and social assistance provisions was illegal in Britain, which prevented policymakers undertaking policy experiments based on either random assignment or area controls. (The 1996 legislative change was introduced specifically to facilitate policy research and evaluation.) In contrast, in the US, experimentation, certainly in the area of welfare, became a mechanism in the struggle by states for increased autonomy and resources. The 1988 Family Support Act allowed states to vary federally funded policies (through the granting of "waivers") only if policy innovation was evaluated by means of random assignment methodology. It is also probably true to say that policy evaluation in Britain is less dominated by positivist economics than in the US, and that sociology and that the policy sciences are more influenced by constructivist and interpretativist traditions. There was therefore less demand for experimentation and greater acceptance of the value of formative evaluation.

A commitment to evidence-based policy

The election of a Labour government in 1997 ratcheted up the pace of change. It was committed to modernising policy and policymaking, to evidence-based policy, and to piloting ahead of implementation. It also advocated increased accountability of government based on publishing indices of policy performance against preset policy targets (Walker, 2001; Cm., 2002). In government, Labour published policy a White Paper "*Modernising government*" (Cm., 1999) which promoted outcome-focussed policymaking, along with a prescriptive handbook "Professional policy making" (Cabinet Office, 1999) that emphasised the use of evidence and argued for "building systematic evaluation of early outcomes into the policy process". A Performance and Innovation Unit was established within Cabinet Office which in turn produced another handbook which attempted to place good analysis at the heart of policy making (Cabinet Office, 2000; Davies *et al* ., 2000).

The centralised nature of the British state ensured that the goal of commissioning evidence for policy could readily be achieved. Around 70 policy pilots – programme evaluations and prototypes – were initiated by central government departments between 1997 and 2002, leaving aside other kinds of officially commissioned policy related research.[4] For the most part, these evaluations were undertaken by independent research organisations, were well funded and used the best designs possible given the constraints discussed below. Moreover, policymakers were hungry for the results from the evaluations that they had commissioned and the findings certainly reached the desks of decision makers. The evaluation results were also mostly published and some reports attracted media comment. There seems no doubt that on many occasions policies were shaped in light of the evaluation findings (Davies *et al.*, 2000).

To take one example, the New Deal for the Long-Term Unemployed implemented nationally in June 1998. This activation measure was less elaborate and less flexible than others introduced by the Labour government and early administrative monitoring demonstrated that it was proving to be less effective in helping people to find jobs (based on gross figures not accounting for substitution and deadweight effects). By November 1998, pilots had been launched in 28 areas to test alternative strategies and these revealed that reducing the waiting time for access to the programme, adding an intensive activity phase and encouraging greater flexibility in approach together increased the numbers leaving benefit for employment by 73 per cent (again only gross estimates were used; Lissenburgh, 2001). In the light of this evidence, the national scheme was remodelled. Unemployed people were admitted after 18 rather than 24 months, compulsory periods of intensive activity modelled on other programmes were added, and flexible packages of intensive support were introduced that included work experience and placements, occupational training and soft skills development.

Political visibility

However, several features of the policy environment conspired against both the effective use of policy evaluation and the most efficient designs. First, Britain's political system is very adversarial and the results of policy evaluations constitute ammunition in the political debate. Many of the policies promoted by Labour ahead of their 1997 election were politically controversial, and ministers felt the need to justify implementation decisions with reference to evaluation evidence and to defend themselves against attacks from opposition groups. As early as October 1997 – just two months after the start of a two-year evaluation of New Deal for Lone Parents (a work activation measure directed towards lone parents receiving social assistance) – the Secretary of State for Social Security was reporting that one in four

participating lone parents were finding work (DSS, 1997). Likewise, opposition spokespersons culled evaluation reports for the most negative findings.

While it is not new for research evidence to be used to justify *a priori* decisions or those taken on exclusively ideological grounds (Greenberg *et al.*, 1999; Walker, 1987), the risk is that such selective reading brings evaluation research, along with "statistics", into public disrepute. A further possibility is that ministers will decide that the political aggravation generated by evaluation outweighs its value as a policy tool.

Timing

Secondly, the pace of policy making during the 1997 Labour administration was exceptionally frenetic, and ministers frequently proceeded with full or extended implementation long before the end of piloting. This had always been intended with certain policies, such as New Deal for Disabled People (an activation measure aimed at recipients of incapacity benefits) when decisions were due to be taken 12 months into a two year evaluation (Loumidis *et al.*, 2001). However, in other cases, new political considerations took precedence over waiting for results. The New Deal for Lone Parents provides an example with national implementation being brought forward in the context of ferocious opposition to government proposals to reduce the benefit paid to lone parents (Hales *et al.*, 2000). To the extent that early results are seldom typical, policymakers who succumb to the temptation to rely on interim findings risk making significant mistakes.

Timetables of research and policy are seldom coincident [although research is likely to have most impact when results become available when policymakers are seeking information (Berthoud, 1984; Walker, 2000b)]. However, in Britain, this incompatibility is most marked with respect to prospective evaluation. The length of parliaments and the average two-year tenure of ministers – a brief period during which career politicians need to demonstrate achievement – both conspire against the lengthy policy experiments common in the US.

Many of New Labour's policy pilots were prototypes, rather than programme evaluations, with various forms of early monitoring being used to fine-tune policy implementation and, to a lesser extent, policy design. Moreover, key policy decisions were often informed by elementary monitoring and by the early results from process evaluations. In this context, it is not self evident that the scale of the pilots and the panoply of evaluation methods and "controls" used in an attempt to secure scientific rigour were warranted.

Power of politicians

Thirdly, in Britain the civil servant is in post to serve their minister (and, through them, the monarch). This gives ministers great authority and, when

they are interested, great influence on the shape of policy evaluations, sometimes down to the minutiae of design. It also places great responsibility on civil servants to argue the case for robust design but means that they have limited authority to insist and there are numerous examples, even since 1997, of political considerations and fears leading to sub-optimal designs. The rejection of random assignment on a number of occasions is a case in point (Walker, 2001).

Complex government

Fourthly, while Britain lacks the complexities of federal government, policy problems typically extend beyond the remit of single departments and many agencies are often involved in delivery. This means that the purpose and mode of evaluation have to be agreed between departments, making for complex and sometimes over-elaborate research, with time consuming negotiation over design of research instruments, etc., often telescoping the time available for fieldwork and analysis.

Hyperactivity

Fifthly, the breadth of Labour's manifesto for change meant that most areas of policy were in a state of flux between 1997 and 2001. At the most basic level, the government's enthusiasm for piloting policies and for geographical controls rapidly exhausted the supply of localities that were not either in use as action areas or controls. More profoundly, the wide sweep of policy objectives limited the number of independent control variables available. Most of the new welfare policies had explicit macro economic objectives as well as individual behavioural ones. This made it difficult to control for exogenous changes, in, for example, economic activity, when the policy in question was intended to operate on a hierarchy of levels (Anderton, Riley and Young, 1999b).

Policy amnesia

Finally, if the adversarial nature of British politics leads to short-termism, it also promotes policy amnesia. It is a British tradition that incoming governments do not have access to the policy files created under their predecessors and, while civil servants can provide some form of institutional memory, the commonly practiced mode of policymaking is to start again from first principles. This approach militates against the systematic accumulation of policy intelligence and evaluation data, and promotes acceptance of the view that research concerning previous policies is of limited relevance and value. While most policy development begins with a trawl of existing evidence, these trawls are better likened to voyages of discovery than to information recall.

The British experience suggests that even with a commitment to policy evaluation emanating from the top of government, the policy environment is not naturally supportive of evidence-based policymaking. While the 1997 election of a Labour government has resulted in policy evaluation, directly commissioned and published by government, becoming ubiquitous policy tool, the policy environment still frustrates the most effective design and use of evaluation. Some of the reasons for this, such as the policy amnesia reinforced by constitutional precedence, are perhaps parochial. Others, including political interference and the discongruence between research and policy timetables, have more universal applicability.

Policy characteristics

Classic experimentation is premised on measuring the effect on one variable – the policy objective – of manipulating another – the policy instrument – while all other variables are held constant. In reality, policies often have a multiplicity of objectives that may be contested, differ between government departments and ministers and change over time. When there are multiple objectives, and hence multiple outcome variables, trade-offs in the effectiveness of the evaluation design are almost inevitable. As already noted, design issues become even more complex when some objectives relate to changes in individual behaviour and others to aggregate effects. Since not all outcomes can be measured with equal precision, there is a need to prioritise policy objectives, at least for the purposes of the evaluation; this is something that politicians are not always willing to do which then results in inefficient designs.

This problem is, of course, exacerbated when the impact of policy reform on any of the objectives is unlikely to be great. Moreover, this is likely to be the norm rather than the exception. White (2000), for example, reviewing welfare to work schemes across Europe found that "most programme impacts fall within plus or minus five points from a 10 percentage point gain in employment" and similarly modest effects have been reported from the vast majority of random assignment evaluations of welfare reforms in the USA (Ashworth *et al.*, 2002b). If small effects are all that can be expected, this has important implications for the size and style of evaluations. It can mean very large sample sizes, sometimes making interview surveys prohibitively expensive and requiring increased or total reliance on administrative statistics. In Britain, administrative files often lack important contextual information – since the law allows only information pertinent to determining a person's entitlement to benefit from a policy to be collected and retained – and primary legislation would be required to allow the quality and relevance of such information to be enhanced. An alternative approach is to increase the

precision of the policy evaluation through design rather than scale, but the opportunities are limited whenever multiple policy objectives are important.

Efficient design not only requires objectives to be prioritised but also prior specification of the degree of change that would constitute policy success. This has rarely been done in the Britain.[5] If the amount of change intended is not specified a priori, it is possible after the event to declare almost any change observed in the preferred direction to be a success. Certainly, dissemination of the results of policy evaluation in Britain has been subject to positive spin, although the published evaluation reports are detailed and appear to be scrupulously fair.

Policymakers designing pilots in Britain have been prone to experiment with a range of policy permutations. Each permutation added to an experimental design either reduces the effective sample size or requires a larger, and hence more expensive, pilot. This problem is exacerbated by the understandable desire to test policies in contrasting environments, which means that each variant needs to be implemented in each kind of locality on a scale that is sufficiently large to enable the effectiveness of the policy permutation to be determined. The evaluation of the Education Maintenance Allowance, a subsidy to encourage young people to remain in school, began with four policy variants in nine areas to which several new variants were later added (Ashworth *et al.,* 2002a). In reality, though, it is usually very difficult to determine whether variation in the outcomes observed between areas is due to the local context or to the process of implementation. Process studies have helped, but much of the between area variation encountered in British evaluations has been left unexplained (Loumidis *et al.,* 2001).

Some of the policies evaluated in Britain have been poorly specified in advance, in part because of a commitment to shape policies according to expertise available in the field. Many of the "New Deal" labour market activation policies have devolved much of the detailed policy formulation to the level of personal advisers dealing with jobseekers or to the partnership organisations involved in implementation. In such circumstances, it has often quite proved difficult to determine what, or which, policy was being evaluated, while treating the package as a whole has left policymakers unclear as to which components of the policy were important in generating results and which were not.

Policymakers have also sought to modify the policy in the light of results from the early stages of the pilot implementation, an understandable but very difficult desire for evaluators to accommodate. In practice, of course, pragmatic policy changes are often made during evaluation, since it is irrational to proceed with an ineffectual scheme. Summative evaluation requires that a uniform policy to be uniformly implemented throughout the

study period: if a change occurs that improved effectiveness, the true impact of the policy will be understated, while if the change has no effect this cannot be formally detected. Policy change can therefore generally only be accommodated in summative evaluations by recommencing the evaluation, or by extending the monitoring period to generate sufficient cases to compare outcomes before and after the modification. In Britain, the evaluation of New Deal for Disabled People was effectively restarted once it had been recognised that personal advisers did not sufficiently prioritise employment-related outcomes (Loumidis *et al.*, 2001). Process studies may help resolve this kind of conundrum by offering insight into how change might have been effected.

To summarise, the attempt since 1997 to build policy evaluation into the heart of practical, policy making in Britain has required the evaluation of complex, poorly specified and unstable policies. In effect, the realities of policy have taken precedence over the requirements of effective evaluation. Without benefiting from the power of random assignment, clear policy goals or stable implementation, estimates of effect size have not been robust and findings on policy impact have generally been equivocal. This has resulted in frustration both for policymakers, who are required to use value for money arguments in their negotiations for resources, and for those charged with the task of drawing up detailed policy designs and regulations (Walker, 2001). It has also led to an increased reliance on formative evaluation in mixed-method designs since this usually generates evidence even in the absence of measurable effects. [There are reports from the US that state officials place greater emphasis on the process elements than impact estimates in deciding on policy options, Greenberg *et al.* (2000)].

Commissioning evaluations[6]

The process of research procurement in Britain significantly shapes the design of policy evaluations. Contracts are typically let competitively with selected organisations being invited to respond to a variably detailed research specification over a period of about four weeks. Potential contractors are usually encouraged to meet with departmental policymakers and researchers in order to fine-tune their understanding of the requirements of the evaluation before submitting a tender. Tenders received are reviewed, often sent to external referees, and one or more organisations invited to present their proposals orally. Post tender negotiations usually follow, sometimes involving more than one of the bidders, to revise designs (sometimes adopting ideas proposed by unsuccessful contractors) and to fix the final programme of work and contract price.

The research specification issued to potential contractor is often very detailed. This reflects the aspirations of government to commission a product that is tailored to their needs, but serves significantly to limit the scope for

creative design. Typically, for example, pilot and control areas will already have been selected on a mixture of research based, pragmatic and political criteria ahead of tendering. The result is to fix not only the location of the evaluations but also the evaluative model (for example, quasi-experimentation with area-based controls). Time scales will also have been fixed, driven by political and implementation priorities rather than by design considerations.

The invitation to tender typically invites alternative designs but this can be a high-risk strategy when viewed from the perspective of a potential contractor. When more than one government agency has been involved in designing the specification, probably involving protracted negotiation, contractors may rightly conclude that the chances of all the agencies accepting a radical departure from the core design are low. Indeed, one of the arts of winning tenders is to place oneself ahead of the competition and flattering departmental staff and following the core design is one way of doing so.

Likewise, the rules of the game are that all parties recognise that time horizons allowed for the evaluation are far too short for all but the most immediate effects of the policy innovation to be assessed. Research contractors typically point out the deficiency but, keen to engage in interesting and sometimes lucrative work, generally take the view that any evaluation is better than none.

The reality is often even worse than the theory. The imperative to deliver results in time to inform policy decisions often means that delays in the approval of research instruments or in fieldwork are not reflected in the reporting schedule. The result is to telescope analysis, risking error and guaranteeing less than thorough investigation of the issues. Four to eight weeks to analyse the results of a 50-minute interview with a complex national sample is quite common.

It is also worth noting that few research organisations have the capacity to conduct large-scale evaluations without additional assistance, especially since work is usually expected to begin as soon as the contract is let (and often before the contract is signed). This means that much of the short tendering period is spent building consortia and determining internal management and financial structures. Indeed, while a competitive tender may easily take 20 person days to compile, this time is often crammed into one or two weeks of the tender period – not necessarily the ideal way to devise an effective evaluative design. The speed of the tendering process also generally serves to preclude most academics whose other commitments typically prevent them responding with sufficient rapidity.

Development of the evaluation design is nested within the policy process and affected by all the constraints noted above. Unlike demonstration projects in the USA, where the importance of developing a robust evaluative design

generally takes precedence over most other considerations, design issues take second place to policy concerns and are typically addressed only quite late in the day. Hence, researchers – both those inside the government machine and potential contractors – are typically presented with the task of devising an evaluation for particular pilot implementation, rather than being asked to determine how best to evaluate a policy. A proposed new evaluation of policies to promote retention and advancement in employment is an important exception and, intended to be Britain's first demonstration project, was deliberately developed in Cabinet Office, outside the main policy department.

This last development is illustrative of recognition within the UK government that the nature of the policy process and the strictures of competitive tendering are inimical to effective evaluation. One response is a revised code of practice relating to the commissioning of research contracts that is currently being drawn up by the Social Research Association (a professional body representing the research community) with the active involvement of government officials. However, political realities can easily un-rail the best of intentions.

The knowledge market place

Britain's centralised government creates a monopolistic purchaser in the market for policy evaluation (monopsony). This reduces the number and potentially the independence of suppliers. It may also contribute to perpetuating the ill-informed public policy debate conducted in the press and other media.

Local government generally lacks the resources to commission large-scale policy evaluations. Also, unlike in the US, the comparatively few research trusts in Britain normally avoid direct involvement in policy evaluation, being more eager to fund research exploring the nature of social problems; there is perhaps an expectation that it is for government to develop and test policy options. As a result, until recently, the demand for large-scale policy evaluation was restricted and the number of suppliers, the level of expertise and the depth of experience were all correspondingly limited. With the escalation in demand since 1997, central government has looked to US expertise to fill gaps in capacity.

For a number of reasons the involvement of academics in policy evaluation has been limited. Over a long period, as a minister admitted in 2000, there was "a seam of anti-intellectualism running through government both at the political level and amongst officials" which had "served to alienate academia" (Blunkett, 2000, p. 16). Certainly, there is little explicit use of social science theory in British evaluation studies, with policymakers inclined to view theory as jargon and preferring to deal with

"facts". Moreover, in return, British academia has tended to view applied research as having less esteem than basic research and policy research for government as having even less. This lack of enthusiasm to engage with government research may also reflect concern about the way in which scholarship may be circumscribed by the political and policy considerations discussed above, and fear of political interference in the dissemination of results. In addition, substantial parts of British social science have reacted against both positivistic and empiricist styles of research and, until very recently, training in research methods, especially quantitative ones, has been poor by international standards. Furthermore, the process of competitive tendering with rapid turnaround has tended to favour full-time research staff working in dedicated applied research units over mainstream academics.

If the demand for evidence based research continues, there is the real danger of a further separation between applied and discipline based social science. The burgeoning evaluation industry may take on more of the characteristics of market research with the connection with academe being restricted to the recruitment of junior staff by applied research institutes.

The monopsonist nature of the British evaluation market may also influence the nature of the political debate about findings from evaluation studies. There are several potential audiences for policy evaluation in addition to policymakers. They include not only opposition politicians but other interest groups and, of course, managers and staff in government agencies and partnership organisations involved in the implementation of policy. The academic and research community should also be interested as, via the media, should be the electorate that indirectly funds the evaluations and stands to be the major beneficiary if the research is well conducted. Each audience is likely to have different concerns, be interested in diverse aspects of the evaluation, and require information to be disseminated in varying forms. However, the prescriptive nature of commissioning research includes the mode and timing, if not the content, of reporting, and is carefully tailored to the needs of central government policymakers. While formal publication can now be expected, except in rare cases where the research is judged not to attain a minimum standard, it usually takes place several months after the written report is delivered and discussed with policymakers. Reports tend to lengthy and not easily accessible by a lay audience; press releases are often issued but media coverage is generally limited. Only rarely are different versions of a report targeted at different audiences and without promotion, reports become documents of record rather than agents of change.

Finally, it is important to reiterate the lack of cumulative learning in Britain. The datasets generated by policy evaluations have generally been archived (after anonymisation) with guaranteed public access, but comparatively few have been subjected to secondary analysis. This may

simply be due to ignorance or lack of funding. More likely, it reflects a lack of academic interest in policy issues and failure by the policy community to appreciate the importance of basic research and theory development in analysing policy issues and devising effective responses. Given the limited time that is often allowed for primary analysis, this constitutes a major under-utilisation of resources.

To summarise, Britain, especially since 1997, has sought to integrate evidence and policy evaluation into the policy process. At the level of activity, it has largely succeeded: policy makers expect that policy will be evaluated and in many case it is. However, whether current practice is sustainable must be open to serious doubt. Results to date, especially in terms of establishing impact and cost-effectiveness, have been disappointing and policy structures and policymaking have accommodated little to the requirements of effective evaluation. A single model of evaluation – loosely conceived of as a prototype – dominates at the expense of other strategies that might better accommodate some of the requirements of policy. Strategic thinking, both with respect to policy development and information requirements, is rare with the result that evaluation is usually undertaken with counterproductive haste.[7] The evaluation community in Britain is limited in size and experience and is employed largely at the behest of central government. While there is little if any evidence of evaluators colluding with government propaganda, many academics have shied away from active engagement in either policy analysis or policy debate. As a result, the quality of evaluation and policymaking and the level of political discourse have all suffered.

Constructing a culture of policy evaluation

The attempt by the United Kingdom to create a bias in the policymaking process towards evaluation is instructive if not reassuring. It is important to recognise, though, that the objective of critique is to identify scope for improvement, which can tend to exaggerate the negative.[8] It is therefore important in this closing section to accentuate the positive message before reflecting on some of the pre-requisites for encouraging the more effective use of improved evaluation evidence.

Policy evaluation could inform all stages in the policy process from the question "what is the problem?" to "how well did we do?" Increasingly often, it does. To the extent that the evaluation is sound, policymaking is likely to be improved and resources better deployed to the collective benefit of all. Moreover, taking Britain as a case study, it is evident that with political commitment, it is possible quite quickly to place evidence at the heart of policymaking and to make heavy use of evaluation research.

What is less clear from the case study is whether the new British culture of policy evaluation is yet secure and whether it is able to deliver improvements in the quality, as well as in the quantity, of evaluative evidence and to ensure that the most robust evidence is used. More generally, it has to be asked whether the principles that underlie good evaluation are compatible with the realities of the policy world and whether a sufficient degree of convergence is possible to create a sustainable new policy paradigm.

Turning to the need for improvement, the preceding review and case study point to five ways in which the development of an evaluation culture might be further promoted. First, there is no single best model of evaluative research. While, for example, experimentation using random assignment may have many attractions, there are numerous occasions when it is inappropriate. However, timing in relation to the policy cycle is always of central importance; evaluation can only gain purchase on the policy process if results are available when required. The challenge is to establish the infrastructure necessary to exploit different evaluative models to address questions relevant at each stage in the policy cycle.

Secondly, to be most effective, evaluative evidence has to be at the heart of the policy process but as distant as possible from the political arena. If is tainted by ideology, or even thought to be tainted by ideology, it loses its power, carrying no more weight than personal opinion, and differing little from propaganda.

Thirdly, real world policies are complex in their objectives and implementation. They are situated in a policy environment that, if anything, is even more complicated. Both the policies and the environment are characterised by rapid and often unpredictable change. All this suggests that evaluators may need always to trade off precision against robustness, and policymakers to accept that evaluation deals only with policy models, simplified representations of policies that share only some of the properties of real thing.

Fourthly, in advanced welfare states the impact of policy innovation is likely to be marginal – the largest effects will probably already have been achieved through the introduction of the public policies that defined the emergence of representational democracy and the welfare state. Almost certainly the effects will be less than those predicted by the advocates of change. This may, on occasion, make summative evaluation prohibitively expensive and, more generally, require reliance on monitoring and more formative approaches.

Finally, policy evaluation is an exceedingly difficult art that calls for substantive knowledge, practical understanding, methodological expertise, and, above all, creativity. There is need systematically to pool expertise and to attract in new talent.

Timing and strategy

The key to developing an evaluation culture is strategy and the key to strategy is timing. With policy horizons so much shorter than those associated with evaluation, it is imperative to establish information streams ahead of the demand for information.

This requires the creation and management of a comprehensive policy information infrastructure. Comprising such an infrastructure would be:

- systematic and continuous scoping of future information needs;
- on-going collection, retrieval and cumulative analysis of administrative information, and monitoring of current policy;
- the development of longitudinal social data and models that facilitate attempts to forecast future policy problems and policy solutions; and
- the construction of comprehensive and accessible databases of evaluative studies and research to facilitate systematic review of evidence.

The approach would aim, as far a possible, to reduce the element of surprise in policy making by engineering longer time horizons. It would place monitoring, systematic review, secondary analysis and meta-analysis at the heart of the evaluation strategy, but these would need to be combined with the accumulation of analysis and theoretical development from a problem, rather than a policy driven, orientation. Making administrative and policy data more readily available to the academic community could stimulate the latter as would proactively and generously funding policy-related, as well as policy-relevant, research.

Information and understanding would be in place to capitalise on the systematic use of policy experience and metaphor to respond to emerging policy issues. This would derive from fine-tuned prospective evaluation using such techniques as micro-simulation, gaming and laboratory experimentation. Policy implementation would typically be accompanied by prototypes or implementation studies designed to describe and understand the ways in which policies were operating. These would benefit from being designed in its own terms rather than, as the British experience, aping impact analyses. On comparatively rare occasions, exploiting the longer time-horizons generated by scooping information needs, programme evaluations could be used to model and test policy principles.

A successful evidence-based policy strategy will also need to stimulate the demand for, and utilisation of, evaluative evidence as well as its production. This requires prioritising evaluative evidence in the policy process by means of regulation, by creating expectations, securing funding and by training policy staff. Policy makers may well require guidance in framing their requests for evidence and in its interpretation and use. This may require

additional training (all senior civil servants in the UK Treasury are now expected to have training in economics) or specialist recruitment of staff with appropriate backgrounds either to work with or to policymakers. Politicians, especially those who are part of the executive, might similarly benefit from tuition on the production and nature of evidence and its potential contribution to the policy process. (In the UK, new ministers receive some induction training of this kind.)

In summary, the reasonable ideal is that the schematic representation in Table 3.1 be transformed into reality, ensuring that the different forms of evaluation are in place ahead of time to inform the various stages of the policy cycle.

Distancing evaluation from politics

The design, execution and initial interpretation of policy evaluations are best undertaken by experts who are not exposed to political pressures. It is therefore important that administrative systems create Chinese walls between politicians and evaluators. In many countries, legislation and practice protect the collection and release of government statistical series from political interference, thereby ensuring that all constituencies can trust the figures they use. The same model should apply to policy evaluation.

Achieving this position may be especially difficult in circumstances where policies are controversial or when politicians have a close personal interest. But, these are the very circumstances where vigilance is most important. The challenge is to achieve distance without reducing the role, effectiveness and influence of evaluation. Placing responsibility exclusively in the hands of independent organisations may make it more difficult to ensure that results reach policymakers when they are most needed, increase the likelihood that the nuances of the policy problem will be missed, and reduce access of evaluators to administrative data for reasons of confidentiality. Some combination of routine procedures for initiating evaluation and receiving findings, together with a clear role for external scrutiny of methods, interpretation and publication may be a more effective strategy.

Independence and freedom from political interference are important for a further ethical reason that is most evident in relation to programme evaluations and prototypes. Unlike ordinary policy research, which is typically passive, programme evaluation seeks explicitly to change peoples' behaviour in order to test the validity of a policy model. Whereas policies implemented by government bodies are generally intended directly to further the collective good, programme evaluation and experimentation achieves this goal only indirectly; the primary purpose is simply to find out what happens when a policy is introduced. Furthermore, implicit in the notion of programme

evaluation is the possibility that things may go wrong. The policy may not work, or not function as well as expected or as effectively as another policy. Evaluation is, therefore, an inherently risky exercise that may potentially do "harm" those participating in the experiment; they may lose income, for example, or suffer "psychological" consequences such as stress and loss of self-esteem (Newman and Brown, 1996).

"The ethical argument in favour of experimentation is that it is preferable to inflict possible harm on a small scale in an experimental study rather than unwittingly inflict harm on a much larger scale as a result of misguided policy" (Burtless, 1995). In these situations, researchers and policymakers must weigh the possible negative consequences of the evaluation against the gains that are anticipated. These decisions need to be taken with reference to the common good, perhaps formally by an independent ethics committee. Certainly, they should not be captive to political expediency.

The ethical premises supporting policy evaluation also have implications for the utilisation of results. An evaluation must be socially important to justify putting people at risk of harm, which in turn requires that full and appropriate attention will be paid to the results in informing policy development. To dismiss findings as unhelpful undermines the legitimacy of the evaluative model.

Complexity and reality

The complexities of policies, together with the expectation of small-scale effects, have important implications for both policymakers and evaluators. They limit the precision with which the outcomes of policy can be measured and may well mean that performance against certain objectives, sometimes important ones, cannot be assessed quantitatively. This means that policy evaluation will never remove all the risks associated with policymaking and that evaluators are unlikely ever to inherit Plato's mantel as philosopher rulers. Rather evaluators will continue to supply evidence for the policy process (hopefully at more frequently and appropriate times) that will continue to be judged and used (albeit by policymakers who are better informed about the evaluation process) alongside other sources of information.

The same considerations of complexity and difficulty of measurement also reinforce the trend towards multi-method evaluations. It may also call for a reassessment of priority given to quantitative measurement and to qualitative understanding, since there will usually be necessary trade-offs between the two. (In Britain, priority has usually been given to measurement but with sometimes disappointing consequences.)

As importantly, the different methods need to be effectively integrated within one design. This often does not happen, not least because the

integration of quantitative and qualitative methods is extremely difficult to achieve (Ritchie and Lewis, 2003). Instead, different methods are implemented in parallel and brought together in a final report rather than being used to define the focus and content of the evaluation itself (Walker, 2001). The result is that interpretations made with reference to the various components of the research retain the status of informed hypotheses rather than anything more substantial. Part of the problem, as noted above, is that the various methods are sometimes drawn from different disciplines; practitioners are separately trained and not equally attuned to the potential and limitations of other methods. In large scale evaluations specialist expertise qualitative and quantitative research can also to be split between different organisations, requiring strong managerial leadership and systems to ensure continuing liaison between all parties.

But, the challenge of integration extends beyond practicalities and often requires negotiation across frontiers between ontological positions. Most researchers engaged in policy evaluation would probably agree that there is a reality to be discovered beyond subjective understanding. However, not all researchers, and probably comparatively few policymakers and politicians, would accept that there are multiple social realities, defined in terms of the vantage points from which reality is observed. Yet, when triangulating between various methods it is often very difficult to reconcile different results if the concept of a single reality is adhered to. This is especially so when results from just two methods are being compared. If they differ it is usually presumed that this is because the results produced by one or other or both methods are wrong, but there is generally no way of distinguishing which. In such circumstances, policy makers are understandably prone to despair and evaluators are tempted to avoid the problem by telling un-joined-up stories based on the different methodologies. There is a pressing need to bring these interpretative difficulties more into the open and to share experience and insight as to ways in which they may be addressed. If it is possible for two apparently conflicting results both to be correct, there needs to be debate as to the circumstances and conditions when this can occur.

Expanding forces

This last observation points to the need to expand the community of scholars and commentators with an interest in policy evaluation. Located at the interface between the policy process and scholarship, the challenges faced by evaluators are currently at the boundary of what social science can offer society.

Yet, from some perspectives in academe, evaluators are bothering themselves with mere puzzles rather than with substantive problems. Certainly, the casual observer of policy evaluation will find more attention given to social science methodology than to social science theory. However,

this is not always the case and when it is, is sometimes explicable in terms of practical considerations. For example, as noted above, in Britain both the policy process and the procedures for commissioning evaluations have discriminated against the involvement of academics.

However, if a culture of evaluation is to be nurtured and sustained, it is essential to connect evaluation technology to its substantive theoretical and methodological roots. This may happen if the strategy for producing evaluative evidence sketched out above can be achieved, creating longer time scales for research, making administrative data more readily available, and accepting that theory has an important role in conceptualising policy problems and in interpreting findings.

Finally, it is also essential to look across national borders since evaluative theory and method (and evaluators) are likely to travel much better than the policies evaluated. The European Evaluation Society already exists and OECD conference for which this contribution was originally drafted marks an important complementary initiative.

Notes

1. A counterfactual could theoretically be constructed by removing the right to access a policy on a random basis but this is likely to raise ethical and political concerns. Even so, policy monitoring which seeks to address the first question and retrospective evaluation, designed to answer the second, are in practice far more common than programme evaluation.
2. Experimentation has been used in evaluation to explore decision taking by frontline administrators (Huby and Dix, 1992).
3. These are often termed compensation questions or options.
4. Definitional problems make a precise number impossible. Adding pilots commissioned in Scotland would probably take the total to 100 or more.
5. It is important to acknowledge that the Labour government in Britain has established targets for broad areas of policy and that the performance of individual government departments is assessed against these targets. The Government is committed to eradicating childhood poverty by2019 although no definition of poverty has been agreed.
6. This section draws heavily on Walker (2001).
7. The Centre for Management and Policy Studies in Cabinet Office was given a strategic cross-government role to address some of these issues but has recently been the subject of reorganisation and downsizing.
8. Indeed, the charge of negativity is one that policymakers often level against policy researchers when they are often only doing their job.

References

ASHWORTH, K. *et al.* (2002a), *Educational Maintenance Allowance: The first two years, a quantitative analysis*, London: Department for Education and Skills, Research Report 352

ASHWORTH, K., CEBULLA, A., GREENBERG, D. and WALKER, R. (2002b), "Meta-Evaluation: Discovering what works best in welfare provision", Paper presented at the Annual Conference of the United Kingdom Evaluation Society, at the South Bank Centre, London, 12 December 2002

BERTHOUD, R. (1984), *The Reform of Supplementary Benefit*, London: Policy Studies Institute.

BLONDEL, S., LOHÉAC, Y. and RINAUDO, S. (2000), "Rational Decisions of Drug Users: an experimental approach", preliminary report, The 20th Arne Ryde Symposium on Experimental Economics.

BREWER, M., CLARK, T., and MYCH, M. (2001), *Credit Where It's Due: An assessment of the new tax credits*, London: Institute for Fiscal Studies.

BRICE, G.C., GOREY, K.M., HALL, R.M. and ANGELINO, S., (1996), "The STAYWELL Program – Maximising Elders' Capacity for Independent Living Through Health Promotion and Disease Prevention Activities", *Research on Ageing*, 18, 2, 202-218.

BULMER, M. (1986), *Social Science and Policy*. London: Allen and Unwin.

BURTLESS, G., (1995), "The Case for Randomised Field Trials in Economic and Policy Research", *Journal of Economic Perspectives*, 9, 2,63-84.

BURTLESS, G. and ORR, L. (1986), "Are Classical Experiments Needed for Manpower Policy?", *Journal of Human Resources*, 21, 4, 606-639.

BOTTOMLEY, D. and WALKER, R. (1996), "Experimental Methods for Policy Evaluation", Loughborough: Centre for Research in Social Policy, Working Paper 276S.

CARTER, N. and GREER, P. (1993), "Evaluating agencies: Next steps and performance indicators", *Public Administration*, 71, (4), pp. 407-16.

CABINET OFFICE (1999), *Professional Policy Making for the Twenty First Century*, London: Cabinet Office: Strategic Policy Making Team.

CABINET OFFICE (2000), *Adding It Up: Improving analysis and modelling in central government*, London: Cabinet Office, Performance and Innovation Unit.

CM. (1999), *Modernising Government*, London: Command Paper, Cm. 4310.

CM. (2002), *Opportunity For All*, Fourth Annual Report, London: Cm. 5598.

CMND 9518, (1985), *Reform of Social Security: Programme for Change*, London: HMSO.

CMND 9691, (1985), *Reform of Social Security: Programme for Action*, London: HMSO.

CHEN, T. (1990), *Theory Driven Evaluation*, New York: Sage.

CULLEN, J. and HILLS, D. (1996), *The Role of Randomised Controlled Trials in Assessing Effectiveness of Services: A Critical Review*, London: The Evaluation Development Review Unit, The Tavistock Institute.

CURINGTON, W.P. (1994), "Compensation for Permanent Impairment and the Duration of Work Absence", *The Journal of Human Resources*, 29, 888-910.

DAVIES, H., NUTLEY, S., and SMITH P. (eds) (2000), *Evidence and Public Policy*, Bristol: Policy Press, September.

ECKEL, C. and GROSSMAN, P. (2001), "Differences in the Economic Decisions of Man and Women: Experimental Evidence". In C. Plott and V. Smith (eds.) *Handbook of Experimental Results*, New York: Elsevier.

GLASS G. (1976), "Primary, Secondary and Meta-Analysis of Research."*Education Research*; 5:3-8.

GIBSON, C. and DUNCAN, G. (2002), "Lessons Learned: The advantages of mixed methods in program evaluation", Paper presented at the American Association of Policy Analysis and Management, Annual Conference, Dallas, 8th November.

GOODYEAR-SMITH F. and FARNELL, A. (2001), "Pain Management in Palliative Care: a modified Delphi consultation study", *www.rnzcgp.org.nz/NZFP/Issues/Apr01/or_gys.htm*.

GREENBERG, D and SCHRODER, M. (1997), *Digest of Social Experiments*, Washington: Urban Institute Press (Second edition, third expected 2003).

GREENBERG, D., MANDELL, M. and WALKER, R. (1999), "Learning from random assignment experiments in the US and Britain". Paper presented at the Association of Public Policy Analysis and Management, Washington, DC, 6th November. Available as Centre for Research in Social Policy, Loughborough University, Working Paper 2264.

GREENBERG, D., MANDELL, M., and ONSTOTT, M. (2000), "The dissemination and utilization of welfare-to-work experiments in state policymaking", *Journal of Policy Analysis and Management*, 19, 3, 367-382.

GUESS, G. and FARNHAM, P. (2000), *Cases in Public Policy Analysis*, Washington: Georgetown University Press.

HALES, J., LESSOF, C., ROTH, W., GLOYER, M., SHAW, A., MILLAR J., BARNES, M., ELIAS, P., HASLUCK, C., MCKNIGHT, A. and GREEN, A. (2000), *Evaluation of the New Deal for Lone Parents: Early Lessons from the Phase One Prototype – Synthesis Report*, London: Department of Social Security (DSS), Research Report 108.

HALL, P. LAND, H. PARKER, R. and WEBB, A. (1975), *Change, Choice and Conflict in Social Policy*, London: Heinemann.

HAMMERSLEY, M. (1992), *What's Wrong with Ethnography*, London: Routledge.

HARRÉ, R. (1986), *Varieties of Realism*, Oxford: Blackwell.

HASLUCK, C., MCKNIGHT, A. and ELIAS, P. (2000), *Evaluation of the New Deal for Lone Parents: Early lessons from the Phase One prototype – Cost-benefit and econometric analyses*. London: DSS Research Report 110.

HILLS, D., WARD, C. BLACKBURN, Y. and YOULL, P. (2001), *Evaluation of the New Deal for Disabled People: Innovative schemes pilot*. London: DSS Research Report 143.

HOOD, C. (1991), "A Public Management for all Seasons", *Public Administration*, 69, (1), pp. 3-19.

HUBY, M. and DIX, G. (1992), *Evaluating the Social Fund*, London: DSS Research Report.

KLEVMARKEN, N. A. (1997), "Behavioral Modelling in Micro Simulation Models", Uppsala: Department of Economics, Working Paper, 31. Accessed from: *www.nek.uu.se/Pdf/1997wp31.pdf*, 28th October 2002.

KNAPP, M. (1984), *The Economics of Social Care*, Houndmills: Macmillan.

LAKATOS, I. (1970), "Falsification and the methodology of scientific research programmes" in I. Lakatos and A. Musgrave (eds.) *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press.

LIPSEY, M. and WILSON, D. (2001), *Practical Meta-Analysis*. Thousand Oaks, CA: Sage.

LISSENBURGH, S. (2001), *New Deal for the Long-term Unemployed: A comparison of provision in pilot and national areas*. Sheffield: Employment Service, Research and Development Report.

LOUMIDIS, J., STAFFORD, B., YOUNGS, R., GREEN, A., ARTHUR, S., LEGARD, R., LESSOF, C., LEWIS, J., WALKER, R., CORDEN, A., THORNTON, P. and SAINSBURY, R., (2001), *Evaluation of the New Deal for Disabled People Personal Adviser Pilot*, London: DSS Research Report 144.

MICHALOPOULOS, S. *et al*. (2002), *Making Work Pay: Final report on the Self-sufficiency Project for Long-term Welfare Recipients, Ottawa*: Social Research and Demonstration Corporation.

NEWMAN, D. and BROWN, R. (1996), *Applied Ethics for Program Evaluation*, Thousand Oaks: Sage Publications.

PATTON, M. (2002), *Qualitative Research and Evaluation Methods*, Thousand Oaks: Sage Publications.

PAWSON, R. and TILLEY, N. (1997), *Realistic Evaluation*, London: Sage.

PLUM, G. (1998), accessed from *www.uni-klobenz.de/~kgt/Learn/Textbook*, 28th October 2002.

ORR, L. (1998), *Social Experiments Evaluating Public Programs With Experimental Methods: Evaluating Public Programs with Experimental Methods*, New York: Sage.

ROGERS, P., HACSI, H., PETROSINO, A. and HUEBNER, T. (2000), *Program Theory in Evaluation: Challenges and opportunities, new directions for evaluation*, 87, San Francisco: Jossey-Bass.

RITCHIE and LEWIS, J. (2003), *Qualitative Research Practice*, London: Sage.

SHAW, I. (2000), *Evaluating Public Programmes: Contexts and issues*, Aldershot: Ashgate.

SMITH, A., YOUNGS, R., ASHWORTH, K., MCKAY, S. and WALKER, R., (2000), *Understanding the Impact of Jobseeker's Allowance*, London: DSS Research Report 111.

STAFFORD, B. GREENBERG, D and DAVIS, A. (2002), *A Literature Review of the Use of Random Assignment Methodology in Evaluations of US Social Policy Programmes*, London: Department of Work and Pensions, In-house Report, 94.

STEWART, J., (1996), "A Dogma of our Times – The Separation of Policy-Making and Implementation", *Public Money and Management*, July-September, 33-40.

THOMAS, P. (1987), "The Use of Social Research: Myths and models", pp. 51-60 in M. Bulmer (ed.) *Social Science Research in Government*, Cambridge University Press.

THORNTON, P. and CORDEN, A. (2002), *Evaluating the Impact of Access to Work: A case study approach*, Sheffield: Research and Development Report WAE138, Claimant Unemployment and Disadvantage Analysis Division, Department for Work and Pensions.

VOYER, J-P., JOHNSON, K., MONTMARQUETTE, C. and ECKEL, C. (2002), *Fostering Adult Education: A laboratory experiment on the efficient use of loans, grants and savings incentives*, Ottawa: Social Research Demonstration Corporation.

WALKER, D. (2000), "You Find the Evidence, We'll Pick the Policies", *The Guardian*, 15th February, p3H.

WALKER, R. (2004), *Social Security and Welfare,* Milton Keynes: Open University Press, Forthcoming.

WALKER, R. (2003), "Applied Qualitative Research" in A. Bryman (ed.) *Encyclopedia of Social Science Methods*, London: Sage.

WALKER, R. (2001), "Great Expectations: Can Social Science Evaluate New Labour's Policies? "*Evaluation,* 7, 3, 305-330.

WALKER, R. (2000a), "Learning if Policy Will Work: The case of the New Deal for Disabled People" *Policy Studies*, 21, 4, 313-345.

WALKER, R. (2000b), "Welfare Policy: Tendering for evidence", pp. 141-166 in H. Davies, S. Nutley and P. Smith (eds.) *Evidence and Public Policy*, Bristol: Policy Press.

WALKER, R. (1997), "Public Policy Evaluation in a Centralised State", *Evaluation*, 3, 5, pp. 261-279.

WALKER, R. (1987), "Perhaps Minister: The messy world of 'in-house' Government research", pp.141-165.in Bulmer, M. (ed.), *Social Science Research in Government*, Cambridge University Press.

WALKER, R. (ed.) (1985a), *Applied Qualitative Analysis*, Aldershot: Gower.

WALKER, R. (1985b), *Housing Benefit: the experience of implementation*, London: Housing Centre Trust.

WALKER, R. with WILLIAMS, J., (1987), "Housing Benefit: Some determinants of administrative performance", *Policy and Politics*, 14, 3, 309-34.

WALKER R., HEDGES, A. and MASSEY, S. (1987), *Housing Benefit: Discussion about reform*, London: Housing Centre Trust.

WEISS, C. (1980), "Knowledge Creep and Decision Creep", *Knowledge: Creation, Diffusion, Utilisation*, 8, 2, 274-81.

WEISS, C. (1992), *Organisations for Policy Analysis: Helping government think*. Newbury Park: Sage Publications.

WEISS, C. (1996), "Theory Based Evaluation: Past, present and future", Paper presented at the American Evaluation Association Conference Atlanta, Georgia 8th November.

WEISS, C. (1998), "Have we learned anything new about the use of evaluation?", *American Journal of Evaluation*, 19, 1, 21-33.

WILENSKY, H. (1997), "Social Science and the Public Agenda: Reflections on the relationship of knowledge to policy in the United States and abroad", *Journal of Health Politics, Policy and Law*, 22, 5, 1241-65.

WHITE, M. and RILEY, R. (2002), *Findings from the Macro evaluation of the New Deal for Young People*, London: Department for Work and Pensions.

WILLIAMS, W. (1998), *Honest Numbers and Democracy: Social policy analysis in the White House, Congress and the Federal agencies*. Washington: Georgetown University Press.

WORTHEN, B. SANDERS, J., FITZPATRICK, J. (1996), *Program Evaluation: Alternative approaches and practical guidelines*, Boston: Addison-Welsey.

YANOW D. (1999), *Conducting Interpretive Policy Analysis*, Thousand Oaks: Sage: Qualitative Research Methods Series, 47.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 4

Evaluating the Impacts of Local Economic Development Policies on Local Economic Outcomes: What has been done and what is doable?

by
Timothy J. Bartik,
Senior Economist The W.E. Upjohn Institute for Employment Research, Kalamazoo, USA

This paper argues that more rigorous evaluations of local economic development policies are feasible. Programs that aid selected small firms can be rigorously evaluated using an experimental approach, without excluding firms from assistance, by randomly assigning some firms to receive more intense marketing efforts by the program. Programs that aid distressed local areas can be rigorously evaluated by random assignment of the program among eligible distressed areas. If an experiment cannot be done, a variety of statistical approaches can be used to compare firms or areas that use the program with comparison groups of firms or areas that do not use the program. These statistical analyses should be supplemented with surveys and focus groups with businesses that use the program, which give some insight into why the program works or doesn't work. Evaluations should go beyond the effects of programs on business growth to effects on local fiscal health and the earnings of the unemployed. Evaluations using rigorous approaches suggest that programs providing information services to small manufacturers are frequently effective. Programs targeting distressed areas are ineffective unless great resources are used over a lengthy period.

Foreword

This paper argues that local economic development policies can and should be more rigorously evaluated. The evaluation should attempt to determine the impact of the policy on local economic outcomes – that is, how local economic outcomes differ compared to what would have happened "but for" the policies.

Programs that provide services or financial assistance to small and medium-sized enterprises (SMEs) can be rigorously evaluated using experimental methods. In such a random experiment, the program would be selectively marketed to randomly chosen SMEs, the "treatment" group, while a control group of SMEs would still be eligible for services but would not receive special marketing efforts. The policy's impact on SMEs can be evaluated by comparing economic outcomes and program usage in the treatment and control groups.

Programs that target distressed local areas for assistance, such as enterprise zones, can also be rigorously evaluated using experimental methods. Areas designated for assistance can be randomly chosen among eligible distressed areas, and the scarce available resources can be more

concentrated on these randomly chosen "treatment" areas. The policy's impact on designated areas can be evaluated by comparing economic outcomes in these treatment areas to the eligible distressed areas that were not randomly chosen for assistance.

If experimental data are unavailable or an experiment is infeasible, local economic development programs can and should be evaluated by statistical analyses of economic outcomes in firms or areas using the programs, or more intensively using the programs (the "treatment" group) and economic outcomes in comparison firms or areas (the "comparison" group). There are a variety of well-developed statistical techniques that attempt to determine how much of the differences in economic outcomes between treatment and control groups is attributable to the program.

These statistical analyses should be supplemented with surveys and focus groups targeting the business clients that use economic development programs. Surveys that are independently administered, ensure anonymity, and ask specific questions can provide additional evidence on the effectiveness of the program in affecting business actions. Surveys and focus groups can also give some insight into how and why a program is effective, and suggest how the program can be improved.

Evaluations should seek to go beyond the impact of policies on increasing local business growth to the benefits of the policy for the public. These benefits include the fiscal benefits for government, and increased earnings for the unemployed or underemployed. Fiscal and employment benefits can be estimated using regional econometric models which are combined with special modules that consider the structure of local taxes and government budgets, and the local labor market.

In the United States, these more rigorous evaluation approaches have been extensively – but by no means universally – used by federal, state, and local organisations concerned with economic development. The results of these evaluations suggest that economic development programs that provide information, training, and consulting services to small and medium-sized manufacturers are frequently effective in improving local business performance. However, programs that target distressed areas, such as enterprise zones, tend to be ineffective if the services and financial assistance offered are too modest to offset the economic disadvantages of the distressed area; more effective economic development programs for distressed areas, such as the Appalachian Regional Commission, mobilize greater resources over a longer time period. Economic development programs frequently have significant fiscal and employment benefits, however the extent of these benefits varies widely, depending on local conditions. Models can estimate these fiscal and employment benefits if the models incorporate the effects of special local conditions.

Encouraging more rigorous evaluation of local economic development policies probably requires the intervention of higher units of government. These higher units of government should provide funding for evaluations and require evaluations when funding local programs. Such intervention by higher units of government is necessary and appropriate because the benefits of evaluating a particular program go well beyond the organisation running the program, and accrue to all organisations that either run or would consider running similar programs, and to the public.

Introduction

This paper considers the best approaches to evaluating the impacts that local economic development policies have on desirable local economic outcomes.[1] The paper is largely based on my knowledge of state and local economic development policies in the United States, but presumably, similar issues arise in evaluating local economic development policies in other OECD countries.

The paper tries to answer nine questions:

- What are the economic development programs that we are trying to evaluate, and why are they important?
- What type of evaluation of these programs is most needed?
- What biases arise in evaluating these programs?
- Can we effectively use experiments with randomization to evaluate economic development programs?
- Can we use statistical methods to make nonrandom comparison groups truly comparable?
- If a local area has an economic development approach that is truly "unique," can it be evaluated?
- Is there other evidence than statistical comparisons with control or comparison groups that might indicate program impact?
- Can we determine why and how a program has impacts or fails to have impacts?
- Can we determine a program's impacts on ultimate rather than proximate economic objectives?

What are the economic development programs that we are trying to evaluate, and why are they important?

By "local economic development programs," I mean programs that provide assistance to businesses that is more or less customized or targeted to the needs of that type of business, with the immediate goal of increasing

business activity in the local economy. (There are, of course, ultimate economic objectives to be achieved by increasing local business activity, which I will address later.)

There are many ways of classifying such local economic development programs. Table 4.1 provides one classification scheme that classifies programs in a way that will later be shown to be relevant for appropriate evaluation techniques. The first type of local economic development programs are those that provide services or financial assistance to only some eligible firms, usually small and medium-sized enterprises (SMEs), with firms either self-selected for assistance or selected by the programs. Such services or financial assistance may include information or training for the enterprise's managers or workers, or public financial support for the enterprise's startup or expansion.

A second type of program provides financial assistance or services to all firms located in a specified area that has been designated as distressed by some higher level of government that helps finance the program. Examples in the United States include the enterprise zone programs sponsored by many state governments, the "Empowerment Zone" program enacted by the federal government under President Clinton's administration, and the Appalachian Regional Commission started in the 1960s.

Table 4.1. **Classification of local economic development policies**

1. Assistance to selected firms (predominantly to small and medium-sized enterprises)
• Training in how to start-up or manage a business
• Public loans/investments or public support for private loans/investments for business start-ups or expansions
• Information/training on implementing new technology or new management techniques
• Firm- or industry-customized training for new workers
• Information/training on exporting
2. Distressed area assistance (enterprise zones and other programs that are typically designed and designated by higher levels of government)
• Tax breaks in local and higher-level government taxes for firms locating or expanding in the designated area
• Enhanced services or infrastructure in the designated area, whether firm-specific or general
3. Whole area programs (typically targeted at manufacturers or other "export-based" firms; sometimes targeted to particular industries)
• Marketing an area and providing site information to new branch plant prospects
• Providing existing businesses and new businesses with help in resolving government regulatory problems
• Expedited provision of site-specific roads and utilities for new plants or expansions, or previous development of industrial parks
• Tax incentives for new or expanded branch plants or corporate headquarters
• Firm-customized training for new workers as incentive for new corporate facilities or expansions
• Support for networks or clusters of firms in an industry to develop better support services such as training
• Technology or industry twist to any of above programs, for example technology-oriented industrial parks, or tax incentives, or training

A third type of program provides assistance throughout the area sponsoring the program, and to all or almost all firms eligible for assistance, although often firm eligibility guidelines target assistance towards the types of firms that are thought to provide the greatest economic benefits. A common target for such programs are manufacturers or other "export-based" firms that export their product outside the area sponsoring the program, although sometimes programs are more narrowly targeted towards a particular industry, such as some high tech industry. These programs include: marketing an area as a location for new corporate facilities; helping resolve government regulatory problems with new facilities or facility expansions; providing tax breaks, site-specific infrastructure, or customized worker training for new or expanded facilities; and working with networks or clusters of firms in an area to enhance local services or infrastructure.

In the United States, it is estimated that roughly $20-30 billion in state and local government spending or tax expenditures is devoted to such "customized" economic development programs annually, with perhaps another $6 billion annually in support from the federal government.[2] The overwhelming bulk of such resources go to whole area programs, mostly in the form of tax incentives. For example, a recent study of the state of Michigan suggests that, of the $700 million in resources (about $70 per capita) devoted to economic development programs annually, over $600 million is devoted to programs that operate throughout the state for almost all eligible firms, and over three-quarters of this $600 million is in the form of tax breaks, most notably reduced property taxes on new or expanded manufacturing facilities (Bartik, Eisinger, and Erickcek 2003).

However, though $20 or $40 billion in resources is significant enough, the importance of local economic development in the United States goes well beyond this relatively narrow definition of local economic development policy. Such a narrow definition focuses on policies that are clearly customized to individual firms or targeted on particular groups of firms, and excludes many more general state and local policies.[3] In state and local debates over taxes, spending, or regulatory policy, the effects of the policy on the state or local area's economic development is always an important consideration (Peterson 1995, 1981). For example, in recent years, almost three-fourths of all states in the US have shifted their approach of apportioning a multi-state corporation's income among the states to an approach that bases half or more of the formula on the state's share of the corporation's "sales", which often enormously reduces corporate income tax collections for firms that export a sizable share of their product outside the state's boundaries (Mazerov 2001; McLure and Herllerstein 2002). This dramatic change in state business tax policy is usually rationalized as a way to promote the state's economic development. Promoting economic development is also used to rationalize many other

changes in state and local policies: other methods of lowering state or local business taxes; lower personal taxes, particularly those paid by high income individuals; reduced welfare benefit levels; changes in workers' compensation laws or unemployment compensation laws; and changes in environmental or health or safety regulations.

Therefore, evaluating local economic development policies is important, not only because of the billions of dollars of resources involved, but also because economic development activity is clearly one of the most important functions of state and local governments in a federal system. Distinguishing between strong and weak claims for the effects of some proposed policy in providing economic development benefits is clearly crucial in having well-functioning state and local governments.

What type of evaluation of these programs is most needed?

The type of evaluation of local economic development policies that is most needed are estimates of the impact of the policies on desirable local economic outcomes. I will call this "outcome impact" evaluation. Ideally, such an evaluation should include estimates of how outcome impacts will vary with any possible change in the scope, scale, design, or management of these policies, or in other words, that from the evaluation we understand fully how and why the policy has its estimated impacts. In addition, an ideal evaluation would not only tell us the policies' impact on local business activity, which is the proximate goal of local economic development policies, but also the policies' impact on the economic well-being of local residents, the ultimate goal of local economic development policies.

Why is "outcome impact" evaluation needed? Only outcome impact evaluation gives us the information needed if policymakers are to make an informed choice regarding the policy option that will maximize social benefits.

In the United States, a great many reports or studies purport to provide "evaluations" or "performance assessments" of economic development policies, but do nothing of the sort. It has become increasingly common for state and local economic development agencies to produce considerable data on program activities, such as numbers of jobs created by assisted firms. Agency reports sometimes claim that this job creation is a "program impact", which erroneously assumes that none of the economic activity would have occurred "but for" the program assistance. Also, state and local economic development agencies often report data on local economic conditions, such as jobs created during a particular time period or reductions in the unemployment rate. Sometimes these reports claim such improvements in local economic conditions as "program impact", which erroneously assumes that any improvements in the local economy are due to local economic development policies.

For example, a study of "business incubators" in the United States, which provide low-cost space, shared support services, and some consulting help to start-up businesses, claimed that "the business incubation programs studied in this project have stimulated the creation of thousands of new jobs throughout the country" (Molnar *et al*. 1997, p. 12). The study goes on to admit that "some jobs credited to the incubator would have been created even if the incubator did not exist, because a certain number of entrepreneurs will always go into business" (*ibid*., 13). However, the study claims that "it is impossible to know after the fact what a firm would have done without the assistance of its business incubator program. Consequently here, as in most research on the impact of business assistance programs, analysis focuses upon gross, as opposed to net, impact" (*ibid*., 13). In contrast to the claims of this business incubation paper, I argue that we can estimate the net impact of the program by estimating what would have happened, on average, if the program did not exist. Furthermore, I believe that the terminology "gross impacts" is misleading, because such numbers are not necessarily impacts of the program.

To avoid confusion, I should emphasize that data on program activities and local economic conditions is often useful. Program activity data helps in managing programs, and local economic condition data helps in understanding the local economy. These data may even be part of the information that is needed to do a true "outcome evaluation" of local economic development policies, which seeks to identify a cause and effect link between program activities and local economic conditions, and quantitatively estimate its magnitude. By itself, however, data on program activities or local economic conditions do not tell us the impacts of policies on outcomes.

Outcome impact evaluation is often expensive in its demands for more data and expertise in statistics and economic modeling. Because such outcome impact evaluation is expensive, it is not clear that such evaluations need to be performed on each and every program run by each local economic development agency. Individual local economic development agencies are probably best advised to reserve outcome impact evaluation for their most expensive programs, for which the possible gains from better policy choices are the greatest. Higher levels of government may provide a useful service by paying for the evaluation of smaller programs, and ensuring that the results are widely disseminated to the local economic development agencies that use, or might use, similar programs.

What biases arise in evaluating these programs?

The ideal – but impossible – study of a government program would borrow a time machine from H.G. Wells or some other science fiction writer, go back in time and eliminate the program but make no other direct

intervention, and then compare the outcomes in this induced alternative world *without* the program to the outcomes in the original world *with* the program. Absent a time machine, the next best alternative is to find some group of entities that are comparable to the group of entities receiving the effects of the program, but this comparison group has no involvement with the program. For local economic development policies of type 1 (see Table 4.1), in which only a subset of eligible firms receive assistance, the comparison group would consist of firms that do not receive assistance. For local economic development policies of type 2, which target distressed areas, the comparison group would consist of areas that are not officially designated as distressed. For local economic development policies of type 3, which serve all eligible firms in the area sponsoring the program, the comparison group would consist of other areas.

For such comparisons to immediately and easily reveal, without statistical torture, the causal effects of the local economic development policies on local economic outcomes, the comparison group will have to be the same, on average, in observed and unobserved characteristics that affect local economic outcomes. Absent experimental data, which will be discussed later in the paper, the group receiving program assistance will generally differ from the comparison group in ways that affect local economic outcomes. Therefore, the assisted group and the comparison group would be expected to experience different changes in economic outcomes, even if neither group received program assistance. As a result, a simple comparison of the two groups will provide a biased measure of program effects.

What are the likely direction of these biases? For local economic development policies that selectively aid firms (policies of type 1), our intuition is that rapidly growing firms are more apt to self-select into participation in the program, precisely because their growth leads them to be more in need of financial assistance and services. There is some evidence that rapidly growing firms are more likely to use selective firm services provided by local economic development agencies (Jarmin 1999). Furthermore, there is some evidence that firm growth is positively correlated over time (Nexus 1999). Under these conditions, firms that participate in the program would have been likely to grow more rapidly in the future even if they had never participated in the program, which will bias evaluations towards overestimating the positive effects of the program. (Of course, particular local economic development programs may have different biases in their evaluations if the programs select firms for assistance in a different way, or if the change in economic outcomes variable that is examined is not positively correlated over time.)

For local economic development programs that target distressed areas (policies of type 2), these distressed areas – by definition – are likely to have higher levels of economic distress than non-designated areas. (For evidence,

see Bondonio and Engberg 2000; Greenbaum 1998; Greenbaum and Engberg 1998.) Therefore, any study that compares the levels of economic outcomes for targeted areas *versus* some comparison group of areas is likely to be biased towards finding negative effects of the program, as levels of economic outcomes are obviously positively correlated over time, and therefore the targeted areas would have higher levels of distress than their comparison group in the future without the program's intervention. It is not as obvious that changes in economic outcomes will differ between targeted areas and comparison non-targeted areas. In fact, some evidence suggests that, in the United States, the correlation between area designation as an enterprise zone and prior area growth is slight (Bondonio and Engberg 2000). (Again, the bias tendencies in evaluations of a particular program will depend on the targeting rules of the program.)

For local economic development programs that serve all eligible firms throughout the area sponsoring the program (policies of type 3), the bias tendencies in evaluations will depend upon what types of areas are more likely to aggressively pursue economic development. The available evidence suggests that, in the United States, incentives do tend to be somewhat higher in states or cities with higher unemployment and previous slow growth (Fisher and Peters 1998). However, these incentives do no more than offset the generally higher effective basic state and local business taxes that prevail in these high unemployment and slow growth areas, so the effective state and local business tax rate after incentives is not strongly correlated with state and local unemployment rates or employment growth. Therefore, studies of the effects of incentives may be biased towards finding less positive effects of incentives on local economic growth, as state and local areas that heavily use incentives would be more likely to grow slowly even without incentives. On the other hand, studies that look at the effects of basic state and local business tax rates on growth may be biased towards finding more positive effects of lower business taxes, as slow growth states tend to have higher state/local business tax rates (for confirming evidence for the same state over the business cycle, see Reed and Rogers 2000).

Can we effectively use experiments with randomization to evaluate economic development programs?

The best feasible way to avoid bias in estimating the outcome impacts of economic development programs is to experiment with the programs by creating some random process which will help determine which entities (firms or areas) will use the program and which will not. Because the process determining the use of the program is random, we know that the program and treatment groups must be the same, on average, in observed and unobserved variables affecting economic outcomes. Any remaining differences in

economic outcomes between the program and treatment groups are either due to the program, or to random factors affecting economic outcomes for a particular firm or area. With a sufficient sample size, these random factors will average out to zero, and we will be able to precisely estimate the true impact of the program on economic outcomes.

To my knowledge, the only economic development evaluation that has relied on data generated from an experiment using random assignment is a study, sponsored by the US Department of Labor, of the effects of entrepreneurship training for UI recipients (Benus, Wood, and Grover 1994). In this experiment, UI recipients in the states of Massachusetts and Washington were first invited to orientation sessions explaining the entrepreneurship training program. The three per cent of UI recipients who expressed interest in such training after the orientation were then randomly assigned to a treatment group that received such training, and a control group that did not. Forty-nine per cent of the treatment group ended up with some self-employment experience, compared to 28 per cent of the control group, with no sign of a different business failure rate in the two groups. Because the treatment and control group, on average, should be the same in observed and unobserved characteristics, we can be confident that, except for random noise, the 21 per cent differential in self-employment experience is due to the entrepreneurial training program. Note that the usual program practice of claiming credit for all business activity associated with the program would exaggerate the effects of the program more than twofold, claiming credit for all 49 per cent of the treatment group that had self-employment experience. Economic development programs cannot legitimately claim credit for all jobs and other business activity that are assisted by the program, because at least some – perhaps all – of this business activity would likely have occurred even without the program.

Random experimentation methods could readily be used with other local economic development policies that only assist a select group of eligible firms. One concern about such experimentation is a reluctance to exclude some firms from services, which is what is done in classical experiments with the control group. Such exclusion can be avoided if the experimentation takes the form of random selection of firms for targeted marketing of the program. Randomization methods would be used to choose which firms would receive an intensive marketing effort, such as letters, phone calls, and personal visits, informing the firm of the services or financial assistance provided by the economic development program. If this marketing is intensive enough, the result should be some significant difference in usage of the program between firms in the treatment group (the group receiving targeted marketing efforts) and the firms in the control group (the group not receiving targeting marketing efforts). However, no firm in the control group that requested

services would be arbitrarily denied services. The difference in economic outcomes (job growth, productivity growth, etc.) between the treatment and control groups of firms, divided by the difference in program usage between the two groups, provides an estimate of the effects of the program. For example, consider a manufacturing extension program designed to improve firms' productivity growth, and a random experiment that intensively marketed the program to a randomly chosen treatment group of firms. If productivity in the treatment group increased 10 per cent, productivity in the control group increased 5 per cent, and program usage in the treatment group was 35 per cent, *versus* 10 per cent in the control group, then the estimated productivity effect of the program is a 20 per cent improvement in productivity [20 = (10 – 5)/(0.35 – 0.10)].[4] Because the treatment and control groups on average only differ in what random number they were assigned, and thereby whether the program was marketed to them, we can be confident that with sufficient sample size this calculation will reveal the impacts on economic outcomes of the program.

Random experimentation could also be done with economic development programs that target distressed areas. In general, there are more economically distressed local economies than a higher unit of government can afford to target with sufficient resources to realistically help turn around a distressed area's economic fortunes. Furthermore, it is unclear whether, among distressed areas, one should target the most or least distressed: the most distressed areas may need help more, but the least distressed may be easer to affect with the right program. Therefore, any effort by program managers to select target areas among all distressed areas are likely to reflect fairly arbitrary judgments. Finally, in practice it is often the case that higher levels of government use political criteria to select which distressed areas will be designated for assistance. For example, during the Clinton administration, in selecting which areas of the United States would be targeted for an "Empowerment Zone" or "Enterprise Community", the final targeted zones were chosen by political appointees, and did not rigidly follow the ranking developed by a selection panel. Given the inherent arbitrariness and political nature of current procedures for designating distressed areas for assistance, there should be no serious ethical issues for such designation to be done using random assignment. If this were done, the designated areas and the undesignated areas would, on average, be the same in observed and unobserved characteristics and growth prospects, and the difference in economic performance of the two groups would be an unbiased estimate of the effects of the program. For such estimates to be precise enough to be useful, there would have to be a sufficiently large number of randomly chosen designated and undesignated areas so that random factors average out. How large the sample size would have to be depends upon how large a program

effect one is trying to detect, and on how much natural variation there is in the economic performance of distressed areas; standard statistical methodologies allow such issues of sample size to be systematically answered. As a rule of thumb, it seems unlikely that in most cases much could be learned without a sample size of at least 20 in each of the groups, the designated and undesignated distressed areas.

For local economic development policies that serve all eligible firms in the entire area, random experimentation is not possible by definition, as these programs are sponsored by the area, and the area government will not control what programs are adopted by the governments of other areas. As mentioned before, in the United States, whole area programs receive the majority of resources devoted to local economic development policies. Thus, for many economic development policies, random experimentation is infeasible.

Can we use statistical methods to make non-random comparison groups truly comparable?

Because experimentation isn't often done with local economic development policies, and is infeasible with some policies, it is important to explore alternatives. We will often have some data on economic outcomes for firms or areas that use a local economic development program or use a program more intensively (the "treatment" group) and those that do not (the "comparison" group). Because the treatment and comparison groups differ in observed and unobserved variables that affect economic outcomes, a simple comparison of outcomes for the two groups may not reveal true program effects. However, there are a number of statistical techniques that can be used to limit or even eliminate the biases resulting from these differences between the treatment and comparison groups.[5]

This is not the appropriate place to go into all the technical details of the appropriate statistical techniques, but briefly, there are at least five statistical techniques, not necessarily mutually exclusive, that can be used to detect the true effect of a program on some outcome variable when the program users differ in other ways than program use from the nonusers. First, we can simply statistically control for observed variables that affect the economic outcome and might be correlated with program use by including these observed variables in the estimation equation that is used to predict the outcome variable. This approach is most effective in reducing the bias in estimation of program impacts when we have data on as many variables as possible that affect the economic outcome of interest and are correlated with program use. This approach cannot correct for biases that might be caused by unobserved variables that are correlated with both economic outcomes and program use.

This approach also assumes that we know the functional form by which the observed variables affect economic outcomes.

A second approach that is a variant of the first goes under the label of "difference-in-differences" estimation, or "difference-in-differences-in-differences" estimation (DD and DDD for short) (Meyer 1995). Under a DD approach, we compare the difference before and after the program of the differences between users and non-users of the program or policy. Under a DDD approach, if we have reason to think that some types of users are likely to be more affected by the policy than another, we can compare the difference between the likely high impact and low impact groups in the user and non-user group before and after the policy. A DD approach is equivalent to assuming that one can do a good job for controlling for other factors affecting economic outcomes by allowing for effects of the time period, and for whether the entity is in a user or non-user group or a high-impact or low-impact group. The limitation of this approach is that there may be many other variables, both observed and unobserved, that also affect economic outcomes and are correlated with program use. The second approach can be combined with the first approach by adding some of these other observed variables to the estimation equation.

A third approach is matching program users with non-users who are similar in observed characteristics. Recent research has revealed that this matching should focus on finding users and non-users who are as similar as possible in their estimated "propensity score", which is an estimated probability given observed variables that a given entity will use the program (Smith and Todd 2001; Heckman, Ichimura, and Todd 1999).

This propensity score should be estimated using variables that predict program use and have a correlation, independent of program use, with economic outcomes. Variables that predict program use but do *not* independently predict economic outcomes should not be included in the prediction of program use. Such variables, both observed and unobserved, provide the variation in program use that is independent of non-program factors affecting economic outcomes.

The propensity score approach works well if we have data on all the variables that do a good job of predicting program use and are also correlated with non-program factors affecting economic outcomes. This matching approach will not work well if there are many unobserved variables that predict program use and are correlated with economic outcomes. In addition, in many cases there may be no reasonably close matches for some users with non-users, and the estimates from a matching approach hence are only valid as average program effects for the types of program users for which we can find good matches among non-users.

A fourth approach is explicitly modeling selection into the program and how it is correlated with unobserved variables affecting economic outcomes (Murnane, Newstead, and Olsen 1985). This requires the estimation of three equations: one equation explaining economic outcomes for program users, a second equation explaining economic outcomes for non-users, and a third equation explaining whether a given entity is a program user. The estimation of the third equation allows a "selection bias correction" term to be added to each of the first two equations, which – in theory – corrects for the bias caused by unobserved variables that affect economic outcomes and are correlated with program use. This approach assumes that we have accurately specified the variables and functional form that should enter all three equations. In addition, this approach assumes a particular statistical distribution for the unobservable factors (the "error terms") that enter all three equations.

A fifth approach requires finding some "instrumental variable" that predicts program use and is uncorrelated with unobservable variables that affect economic outcomes (Angrist and Krueger 2001). Under this instrumental variable approach, we only examine the change in economic outcomes that can be attributed to shifts in program use that are statistically associated with shifts in the instrumental variable. The intuition is that the effects on economic outcomes of these instrument-induced shifts in program use show the true effects of the program because these shifts in program use will be uncorrelated with unobservable variables predicting economic outcomes, as these shifts are generated by an instrumental variable that is uncorrelated with unobservable variables predicting economic outcomes. The problem is finding such instruments. The instrumental variable must do a good job of explaining program use. Otherwise, the estimation approach throws away too much information. But the variable must have little (ideally, zero) correlation with unobservable variables affecting economic outcomes, and it is difficult to test assumptions about the correlation of a proposed instrument with unobservable variables. Good instruments are hard to find and may not be convincing to all readers.

These five approaches can be combined in different ways. For example, researchers can create matched data sets, include controls for various observed variables in the estimation, and use instrumental variables for the program variable.

There are many good examples of impact outcome evaluations for local economic development policies that use non-experimental data. For programs providing assistance to selected firms, Holzer *et al*. (1993) implicitly used an instrumental variable approach to study a program providing customized training to a firm's workers. This program, run by the state of Michigan, provided grants to manufacturing firms for worker training. The comparison group was firms that applied too late in the fiscal year to receive a grant. The

implicit assumption is that the time a firm applied for a grant is an "instrument" that explains participation in the program, but is uncorrelated with unobservable variables affecting a firm's performance. The study found that firms that received grants had significantly lower scrappage rates after that training was completed than firms that applied for grants but did not receive them. The study would yield biased results if firms that applied late in the fiscal year differed in ways we cannot control (*e.g.,* if such firms were more poorly managed).

Another good study of a firm selective program is Jarmin's evaluation of the federally-sponsored Manufacturing Extension Partnership, which provides consulting advice to small and medium-sized manufacturers in improving their productivity (Jarmin 1999). Jarmin's paper first does a rough match by only including non-clients in the data if they were located in the two states where all his clients were located. The paper then controls for selection bias by estimating an equation predicting whether a given manufacturing firm becomes a client of the MEP (for example, this is affected by whether the firm happens to be in a metropolitan area that has a MEP center), and including a selection bias correction in two equations predicting a firm's productivity growth, one equation for firms that are clients of MEP, and another equation for firms that are not clients of MEP. Jarmin's study finds that MEP increases productivity by 3 to 16 per cent.

For targeted areas, a number of studies in the United States have attempted to evaluate enterprise zones by comparing the performance of enterprise zones to matched non-zone areas. Several studies by researchers at Carnegie-Mellon University (Bondonio and Engberg 2000, Greenbuam 1998, Greenbaum and Engberg 1998) have explicitly made such matches using estimates of the "propensity score", that is estimates of the probability of a given area (in this case, a postal "zipcode" or routing code) being designated as an enterprise zone. These studies find little or no effect of enterprise zone designation on the growth of local business activity. In addition, as mentioned before, the propensity score estimation suggests that enterprise zone designation is not strongly correlated with previous area growth, which increases the odds that the estimates reveal the true effect of enterprise zone designation. Other studies have also examined the performance of enterprise zones with non-zones (*e.g*. Papke 1993, 1994; Hebert *et al*. 2001), but published versions of the research do not contain sufficient information to judge the validity of the matching. Finally, one forthcoming study (Peters and Fisher 2002) evaluates state government-designated enterprise zones by comparing the performance of enterprise zones with high *versus* low levels of incentives. The assumption is that unobservable factors affecting an area's performance might be correlated with the area's designation as a zone, but will not necessarily be correlated with the magnitude of the zone incentives, which

depend on the various political compromises in whatever enterprise zone bill was enacted by that particular state. This study finds little effect of the magnitude of a zone's incentives on firm start-ups or expansions.

For local economic development programs that assist all eligible firms in an area, there is a huge literature on the effects of state and local taxes on business location and growth, which has been summarized by Bartik (1991, 1992) and Wasylenko (1997). These studies typically deal with possible correlations of taxes with other variables affecting business location and growth by including as many relevant location and growth factors as possible as explanatory variables in the estimating equations. As summarized by Bartik and Wasylenko, these studies generally come up with an elasticity of state and local business activity with respect to state and local business taxes in the range from !0.1 to !0.6. That is, a 10 per cent reduction in overall state and local business taxes will eventually increase a state's business activity by 1 to 6 per cent.

Over the last 10 years, Andrew Isserman and his colleagues have done a number of papers that evaluate various economic development interventions in the US by matching counties with the interventions with comparison counties without the interventions, using pre-intervention data. One such study indicated that the Appalachian Regional Commission increased growth in Appalachian counties compared to matched counties outside Appalachia, with this growth effect strongest in counties in which the ARC built highways (Isserman and Rephann 1995). Another study evaluated a large tax cut in the state of Illinois by matching each county in Illinois with similar counties outside Illinois, and found that the tax cut had some short-run economic growth effects but no significant long-run effects (Rogers and Reed, forthcoming).

Finally, one of the best studies of the effects of state taxes compares the business location decisions among US states of foreign firms from two groups of countries: countries in which US state taxes can be credited against the firm's tax liability in its home country; and countries in which US state taxes can only be deducted against taxable income subject to home country taxation (Hines 1996). For firms from the former countries, US state taxes should be irrelevant to business location decisions for any firm with positive tax liabilities in its home country. Hines (1996) found that firms from the first group of countries located in higher tax US states than firms from the second group of countries. This can be seen as a form of DD estimation or instrumental variable estimation. The implicit assumption is that the only difference between the two groups of firms that is relevant to their business location choices is how their home country treats US state taxes. If this assumption holds, then the resulting estimates are convincing evidence that state and local taxes do affect the business location decisions of large corporations in the United States.

If a local area has an economic development approach that is truly "unique", can it be evaluated?

Experimentation or statistical analysis using comparison groups assumes that one has data on a significant number of firms or areas using the program. These statistical methods need a sufficient sample size of program users so that one can assume that unique factors affecting economic outcomes of program users average out over the sample. But what if an area has a unique economic development program? For example, what if the area has some unique package of economic development programs that are believed to have a synergistic effect, so that the entire effect of the package cannot be accurately predicted even if one knows the effects on economic outcomes of each individual program?[6]

If this package of economic development programs is offered to a select group of eligible firms, then firms that do not use the program can be used as a control or comparison program, using either experimental methods or non-experimental statistical analysis, depending upon what evaluation resources and will power are available. The procedures would be identical to what has previously been described. Similarly, if the package is offered to a targeted group of distressed areas, then distressed areas that aren't targeted can be used as a control or comparison group, as described previously.

On the other hand, what if the package is offered to all eligible firms in the entire area? In that case, then the best that any statistical analysis can say is whether the area's economic outcomes differ significantly from the average performance that one would predict, based on the performance of matched comparison areas or based on a prediction equation using characteristics of comparison areas. With only one area offering this program, its economic outcomes will have to differ quite a bit more from the average predicted for other areas for its outcomes to be statistically significantly different, compared to a situation where a sizable number of areas offer the same program. In addition, even if the area's performance is statistically significantly different from what would be predicted, all one can conclude is that the net effects of the area's unique economic development programs, and any other special characteristics of the area during this time period, result in a net effect on economic outcomes that is significantly different. Separating out what is clearly due to the unique program is impossible.

Is there other evidence than statistical comparisons with control or comparison groups that might indicate program impact?

Given the difficulties and uncertainties associated with experiments or statistical comparisons of programs, it is important to consider whether there are alternative methods that can substitute or supplement for experiments or statistical comparisons, and allow us to make some inference of a link between

economic development program activities and economic outcomes. I would argue that there are at least three alternative methods to link program activities and economic outcomes. First, in some cases, if one believes one can model how different programs affect business decisions, then it may be possible to extrapolate from results obtained for other programs to new programs. For example, at current US state and local business tax rates, an elasticity of state business activity with respect to state/local business taxes of 0.25, which is close to the median result in the research literature, implies that it costs roughly $9 000 annually in foregone state and local business tax revenue to create one job, or discounting at a real discount rate of 10 per cent, a present value of foregone state and local business tax revenue of $90 000 (Bartik, Eisinger, and Erickcek 2003; Bartik 1992). If one is willing to assume that all that matters in an economic development subsidy is its cost, and that all cost reductions have roughly similar effects on business location probabilities, then one can infer a likely effect of different business subsidies on economic development. For example, suppose that a new branch plant with 1 000 employees is given economic development subsidies whose present value is $30 million. Then, to be fully consistent with the business tax and location literature, these subsidies would be expected to increase the odds of the branch plant choosing the state by one-third, because this effect on the location probability would yield a present value of cost per job created of $90 000 (= $30 million/[(1/3)(1000)]). Of course, in any particular case, the subsidy was either decisive in tipping the location decision or it wasn't. If one has information that makes it more or less likely that the subsidy was decisive, it should be used. But in the absence of other information, it is unclear why the effects of economic development subsidies on business location probabilities should differ from those of general state and local business taxes.[7]

Second, surveys of firms receiving assistance of economic development programs can, if properly run, be used to get a rough idea of the effects of some economic development programs on business decisions. It has become very common for state and local economic development agencies to use some sort of "customer satisfaction" survey of clients. The more useful surveys, however, ask specific questions about how the assistance provided to the firm has affected its behavior. Responses to such "outcome impact" survey questions are more credible when asked for economic development assistance that is provided in the form of in-kind services rather than cash, because firms have an incentive to claim that cash assistance had an impact to keep the cash coming, whereas it is unclear why a firm would claim an in-kind service was useful if it was actually useless. A good example of economic development surveys to determine the impact of economic development services are the regular surveys of program clients of the Manufacturing Extension Partnership, whose local centers provide assistance to small and medium-sized manufacturers in

improving their productivity. In the most recent MEP surveys, conducted in 2001 for clients whose projects closed a year earlier, about 64 per cent of surveyed clients reported that their involvement with MEP had led to productivity improvements (NIST 2002). The average MEP client reported that the MEP services led to sales increases of $143 000 and cost savings of $50 000. It is unclear why MEP clients would seek to fabricate such responses, particularly since they were provided anonymously to a third-party survey organisation.

The credibility of firms' responses to surveys about the impact of tax breaks and other financial assistance is more questionable. Firms clearly have some incentive to claim the financial assistance affected their location or expansion behavior, in the hopes of keeping the assistance going. In some cases, the program may even have required that the firm sign a stipulation that the subsidy was essential to the location or expansion decision in order to receive a subsidy. However, it is certainly not the case that firms will always claim that financial assistance was crucial. For example, the Colorado state legislative audit agency, in an audit of state's enterprise zone program, surveyed 18 businesses that had located or expanded in Colorado enterprise zones, and found that 10 of these businesses reported that the enterprise zone's incentives had no effect on their location or expansion decision (Hinckley and Hsu 2000). Surveys about financial assistance are more credible when administered anonymously by an independent agency, especially when the surveys ask specific and definitive questions (for example, "did you consider other locations?" or "would the location you chose have been clearly inferior in profitability to these other locations without the subsidy?").

A third method of inferring a link between program activities and local economic outcomes is determining whether administrative data on the program and its clients are consistent with the program's stated purpose. For example, the Capital Access Program in Michigan was designed to encourage banks to provide higher-risk loans to small business borrowers. For each small business loan program, the bank and borrower would each put 1.5 to 3.5 per cent of the loan's value into a loan loss reserve fund, and CAP would provide a 150 per cent match of the bank and borrowers' contribution to the fund. Administrative data suggested that the resulting program had a loss rate of about seven times that of a normal bank loss rate on small business loans, which means that the program – at the very least – is probably encouraging loans that otherwise would not have been made (Rohde, Cash, and Ammerman 1990). This finding is consistent with the hypothesis that the program is expanding the supply of credit to small business, although it doesn't definitely prove that the program would pass a benefit-cost test. As another example, the MEGA program in Michigan provides very large refundable tax credits to a select group of new firms or firm expansions in Michigan, but only if the firm can present financial information showing that without the subsidy, the firm would

locate outside the state. Although it is obviously possible for businesses to make up such financial data, the state is free to ask probing questions about the firm's data analysis, and refuse to provide the MEGA credits if the firm's responses are insufficiently convincing. The requirement that the firm financially demonstrate that the subsidy will be decisive at least increases the difficulties and costs for firms with a relatively weak case of applying to the MEGA program. Officials in Michigan's economic development agency claim that they screen out over 90 per cent of firms expressing an interest in the MEGA program (Bartik, Eisinger, and Erickcek 2003).

Can we determine why and how a program has impacts or fails to have impacts?

One concern about outcome impact evaluations is that they are often perceived, even if done well, as only telling us whether a program works, and leaving the workings of the program a "black box": we don't know why or how the program works, so we don't know how to improve the program. In principle, statistical analysis using control or comparison groups can give insights into why and how a program works if a sufficient variation in program designs is observed and accurately measured. With data on many evaluations, statistical comparison with control or comparison groups can suggest which program designs are most effective, which, in a practical sense, is as good as knowing how or why a program works.

In the real world, however, one rarely observes a sufficient variation in program designs to adequately answer all the important questions about how or why the program works. In particular, it is impossible in principle to have data on program variations that have not yet been tried.

This suggests that surveys of clients and client focus groups may often be valuable in opening the "black box," and getting more insight into the strengths and weaknesses of the program. Statistical analysis using control or comparison groups, and surveys and focus groups, should be seen as complementary approaches to evaluating a local economic development program. The statistical comparisons are more likely to give objective quantitative evidence on the bottom line for the program – its impact on local economic conditions – whereas surveys and focus groups are more apt to give information on how that bottom line can be improved.

Can we determine a program's impacts on ultimate rather than proximate economic objectives?

Most of the discussion so far has not considered what economic outcomes should be evaluated. In practice, the economic impacts that are easiest to evaluate are the proximate impacts on various dimensions of

business activity, such as the number of business start-ups or expansions, job growth, productivity growth, etc. But public subsidies for local economic development programs cannot be justified by these programs' effects on local business activity alone. Public subsidies for local economic development require that the changes in local business activity lead to broader public benefits. The most plausible of such public benefits are fiscal benefits to state and local governments, and employment benefits to local residents.

There are possible efficiency rationales for some programs that promote increased business activity, even if there are no broader public benefits. For example, some local economic development policies can be justified by various "market failures" in information markets or financial markets (Bartik 1990). Private information markets may sometimes fail to provide businesses with information that is more valuable than the cost of providing the information, justifying public provision or public subsidy of the information. Financial markets may sometimes fail to make business loans or investments whose private return exceeds the costs, potentially justifying some public investment or subsidies for private loans or investments. But if all that these public interventions do is promote greater business activity, with no broader public benefits, it is unclear why the public at large should pay for these interventions. The business community at large would be a more justifiable source of funds.

The public receives benefits from local economic development if the increased local business activity leads to fiscal benefits or employment benefits (Bartik 1991). Fiscal benefits occur when the increased business activity, and the spinoff effects of this increased business activity on the local economy, result in tax revenue that exceeds required public expenditure increases. Employment benefits occur when the wages of the newly created local jobs exceed the "opportunity costs" of the non-working time foregone by local residents who obtain jobs because of the newly created jobs. The new jobs must either be filled by employed local residents, non-employed local residents, and in-migrants. Jobs filled by employed local residents lead to vacancies that are filled in this same way, so ultimately the newly created jobs are either filled by non-employed local residents or in-migrants. If these jobs hadn't been created, in-migrants could have moved to another similar local area and obtained a job, so the opportunity cost of their time is close to their wage rate. For local residents who are non-employed, the opportunity cost of their time – their "reservation wage" – may be considerably less than the wage rate of the new jobs.

To calculate fiscal and employment benefits of local economic development policies requires an economic model that takes the initial effects of the policies on local business activity, and calculates the impacts on the overall local economy, including multiplier effects on suppliers and retailers,

and effects on local population growth. A variety of such regional models are commonly used by economic development agencies in the United States, most prominently the REMI model and the IMPLAN model. Once the overall impacts on all local business activity and population growth are determined, these impacts need to be translated into impacts on state and local government budgets, and local employment benefits.[8]

Fiscal impact models that translate economic impacts into budget impacts need to be specially constructed for each state or local area, given variations in the local economy and local tax and budget structure. Fiscal impacts on state and local governments depend on several factors. First, fiscal impacts depend, in part, on how different tax bases are taxed by the state and local governments. Second, fiscal impacts depend on the amount of population in-migration compared to increased business activity, as it is generally the case in the US that the average business pays more in normal taxes than it directly requires in public services, whereas the average household uses more public services than it pays in taxes (Oakland and Testa 2000). Third, fiscal impacts depend, in part, on the share of state and local spending that goes towards purposes related to income redistribution, such as welfare or Medicaid (the United States' medical assistance program for the poor), as such spending will not respond proportionately to local economic growth. Fourth, fiscal impacts depend, in part, on whether the existing state and local infrastructure has some unused capacity that will allow economic expansion without requiring expensive construction of new infrastructure.

In the United States, positive fiscal impacts of local economic development may often be significant. For example, one recent study calculates that positive fiscal impacts of state economic development in Michigan may offset as much as half of the gross costs of the state's economic development subsidies (Bartik, Eisinger, and Erickcek 2003).

Studies of local labor markets in the United States suggests that for every one per cent in extra employment growth in a metropolitan area, local employment rates increase by about 0.2 per cent, and average earnings per job increase due to occupational upgrading by about 0.2, so average earnings per local resident increase by about 0.4 per cent (Bartik 1991). These earnings effects may be long-lasting if the extra employment experience for previously employed local residents increases their job skills, self-confidence, and reputation with employers (Bartik 2001). Other estimates suggest that these positive effects on local earnings are greater if the extra employment growth is concentrated in jobs that pay well relative to the skills required, such as manufacturing jobs (Bartik 1993). Theoretical models of local labor markets suggests that the reservation wages of newly employed local residents are likely to be greater in metropolitan areas with higher unemployment rates, in

which the unemployed are more likely to be desperate to get a job (Bartik 1991). Some attempts to apply these local labor market models in the evaluation of local economic development policies have been made by some state economic development agencies, most notably in New York state (Poole *et al*. 1999). However, we know less than we should about how the employment benefits of local economic development policies are affected by other factors. For example, we might expect that local employment benefits will be greater when there is a better match between the newly created jobs and the job skills of the local unemployed, but there is no direct empirical evidence to support this expectation. We might also expect that local employment benefits will be greater when local labor market institutions are more efficient in job training and job matching, but again there is no direct empirical evidence. More studies of the factors affecting the link between local economic development and local employment benefits would help improve the quality of evaluations of local economic development policies.

Conclusion

This paper has tried to show that more rigorous evaluation of economic development programs is feasible. Such rigorous evaluations can be done through random experimentation, statistical analysis of program users and comparison groups, surveys and focus groups, and linking regional econometric models with fiscal impact and local labor market models.

Such rigorous evaluations have been done extensively in the United States. These studies often find that services to small and medium-sized manufacturers can be effective in improving the performance of these firms. Programs that target distressed areas tend to be ineffective if, like enterprise zones, they provide modest resources, but are more successful if, like the Appalachian Regional Commission, they provide extensive resources over a lengthy time period. Programs that provide financial incentives or services to encourage the economic development of a whole area, such as state tax incentives, have modest effects in increasing employment growth, which is reasonable given the modest share of costs that can be influenced by state or local governments. However, if a state or city has extensive underused infrastructure or labor, then even modest increases in job growth may offer considerable fiscal and employment benefits.

However, rigorous evaluation is still the exception rather than the rule. There are far too many cases where state and local economic development organisations claim credit for any state or local job growth, or at least for any of the growth that the organisations happen to have subsidized.

How can more rigorous evaluation be encouraged? I see little prospects of significant increases in rigorous evaluations without some outside pressure

and funding. Economic development organisations, at least in the United States, face some significant disincentives in rigorously evaluating themselves. In the US context, with suspicion of government activism, negative evaluations are a common excuse for terminating a program. Hence, for US policymakers, the advice of economists Gary Burtless and Robert Haveman often seems reasonable, "If you advocate a particular policy reform or innovation, do not press to have it tested" (Burtless and Haveman 1984, p. 128).

Therefore, I was not surprised a few years ago when one state economic development official told me that the trouble with universities evaluating his programs is that the universities seemed to think that negative evaluations should be made public. One can deplore this attitude, but should understand the real fear of budget cuts and program extinction that motivates it.

Rigorous evaluation of economic development policies is only likely to occur if funded or required by outside groups. These outside groups could include legislatures, governmental audit bureaus, and higher levels of government. It makes sense for these groups to require and pay for evaluations, because the benefits of evaluations largely accrue outside of the agency managing the program. Evaluations of a state or local agency's economic development programs provide benefits to the general public in the local area, who benefit from any improvements that result in government effectiveness, and to local areas elsewhere that have similar programs or might consider similar programs.

Rigorous evaluation is also more likely to occur if the results are more frequently used to improve programs rather than kill the programs. If the basic rationale of the program makes sense, in that the program is addressing some problem that may benefit from government intervention, then negative evaluation results should be used to motivate the creation of a new approach to addressing the problem. This is more likely if the rigorous evaluation is accompanied by data from surveys and focus groups that give some insights into how the program can be improved. A balanced mix of rigorous and "softer" evaluation techniques, and a judicious use of evaluation results, will encourage economic development agencies to be more open to rigorous evaluation approaches.

Notes

1. I have previously considered these issues in Bartik and Bingham (1997). The present paper updates my thinking on this topic and considers more recent research findings.
2. The sources for these estimates are discussed in Bartik (2001, p. 251), and are consistent with more recent estimates in Bartik, Eisinger, and Erickcek (2003). In the United States, unlike in Europe, there is no systematic collection of data on

state and local economic development program budgets and tax expenditures (Thomas 2000). These figures must be extrapolated from individual state studies.

3. For example, Thomas (2000) gets state/local business subsidies in the United States of close to $50 billion annually by including some of the more general state and local tax expenditures for business.

4. This discussion glosses over the issue that the program effect may vary across firms. Technically, all that this experiment can estimate is the effect of the program in the extra 25 per cent of all firms that are induced by marketing to use the program, which may differ from the productivity effect among the 10 per cent of firms which use the program without marketing, or the remaining 65 per cent of firms that are unaffected by marketing. This is sometimes called the "local average treatment effect" (LATE), or the "marginal treatment effect" (Heckman and Vytlacil 2001; Imbrens and Angrist 1994).

5. Much of the following discussion in this paper is phrased as if the program is measured by a zero-one dummy, in which the treatment group uses the program and the comparison group does not. However, the discussion is generalizable to a situation in which there are different levels of use of a program, and the program usage variable is a continuous variable.

6. If there is no special synergistic effect of multiple programs, then the effects of a "unique" combination of programs can be extrapolated from studies of the effects of each individual program.

7. This assumption that all that matters are overall business costs is based on the assumption that output effects dominate in determining business location and expansion behavior, and that factor substitution effects are of secondary importance.

8. For more extensive discussion of state econometric models and associated fiscal impact and employment impact models, and their use in evaluating economic development incentives, see the report by Poole, Erickcek, Iannone, McCrea, and Salem, 1999.

References

ANGRIST, J.D. and A.B. KRUEGER (2001), "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. "*Journal of Economic Perspectives* 15(4): 69-85.

BARTIK, T.J. (1990), "The Market Failure Approach to Regional Economic Development Policy." *Economic Development Quarterly* 4(4): 361-370.

BARTIK, T.J. (with P. EISINGER and G. ERICKCEK) (2003), "Economic Development Policy in Michigan." in *Michigan at the Millennium*, C. Ballard, P. Courant, D. Drake, R. Fisher, and E. Gerber, eds. East Lansing, Michigan: Michigan State University Press.

BARTIK, T.J. and R.D. BINGHAM (1997), "Can Economic Development Programs be Evaluated?" In *Dilemmas of Urban Economic Development*, R.D. Bingham and R. Mier, eds. (pp. 246–277). Thousand Oaks, California: Sage Publications, Inc.

BENUS, J.M., M. WOOD and N. GROVER (1994), "A Comparative Analysis of the Washington and Massachusetts UI Self-Employment Demonstrations". Unpublished report prepared for the US Department of Labor, Employment and

Training Administration, Unemployment Insurance Service under Contract No. 99-8-0803-98-047-01.

BONDONIO, D., and J. ENGBERG (2000), "Enterprise Zones and Local Employment: Evidence from the States' Programs." *Regional Science and Urban Economics* 30: 519-549.

BURTLESS, G., and R.H. HAVEMAN (1984), "Policy Lessons from Three Labor Market Experiments." In *Employment and Training R&D*, R. Thane Robson, ed. Kalamazoo, Michigan: W.E. Upjohn Institute for Employment Research, p. 128.

CCC (1991), *Who Benefits from State and Local Economic Development Policies?* Kalamazoo, Michigan: W.E. Upjohn Institute for Employment Research.

CCC (1992), "The Effects of State and Local Taxes on Economic Development: A Review of Recent Research." *Economic Development Quarterly* 26: 102-110.

CCC (1993), "Economic Development and Black Economic Success." Upjohn Institute Technical Report No. 93-001. Kalamazoo, Michigan: W.E. Upjohn Institute for Employment Research.

CCC (1994), "Tax Policy and Urban Development: Evidence from the Indiana Enterprise Program." *Journal of Public Economics* 54: 37-49.

CCC (1995), *The Price of Federalism*. Washington, DC: The Brookings Institution.

CCC (2001), *Jobs for the Poor: Can Labor Demand Policies Help?* New York and Kalamazoo, Michigan: Russell Sage Foundation and W.E. Upjohn Institute for Employment Research.

FISHER, P.S., and A.H. PETERS (1998), *Industrial Incentives: Competition Among American States and Cities*. Kalamazoo, Michigan: W.E. Upjohn Institute for Employment Research.

GREENBAUM, R.T. (1998), "An Evaluation of State Enterprise Zone Policies: Measuring the Impact on Business Decisions and Housing Market Outcomes." Doctoral dissertation, Carnegie Mellon University, December.

GREENBAUM, R.T. and J.B. ENGBERG (1998), "The Impact of State Urban Enterprise Zones on Business Outcomes." Discussion paper No. 98-20. Washington, DC: Center for Economic Studies, Bureau of the Census.

HEBERT, S., A. VIDAL, G. MILLS, F. JAMES and D. GRUENSTEIN (2001), Interim Assessment of the Empowerment Zones and Enterprise Communities (EZ/EC) Program: A Progress Report. Washington, DC: Office of Policy Development and Research, US Department of Housing and Urban Development.

HECKMAN, J.J., H. ICHIMURA and P. TODD (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64(4): 605-654.

HECKMAN, J.J. and E. VYTLACIL (2001), "Policy-Relevant Treatment Effects." *The American Economic Review* 91(2): 105-111.

HINES, J.R., Jr. (1996), "Altered States: Taxes and the Location of Foreign Direct Investment in America". *American Economic Review* 86(5): 1076-1094.

HINKLEY, S., and F. HSU (2000), "Minding the Candy Store: State Audits of Economic Development." Washington, DC: *Good Jobs First,* September.

HOLZER, H.J., R.N. BLOCK, M. CHEATHAM and J.H. KNOTT (1993), "Are Training Subsidies For Firms Effective? The Michigan Experience." *Industrial and Labor Relations Review* 46 (July): 625-636.

IMBENS, G. and J. ANGRIST (1994), "Identification and Estimation of Local Average Treatment Effects". *Econometrica* 62(2): 467-475.

ISSERMAN, A. and T. REPHANN (1995), "The Economic Effects of the Appalachian Regional Commission: An Empirical Assessment of 26 Years of Regional Development Planning." *Journal of the American Planning Association* 61(3): 345-364.

JARMIN, R.S. (1999), "Evaluating the Impact of Manufacturing Extension on Productivity Growth." *Journal of Policy Analysis and Management* 18(1): 99-119.

MAZEROV, M. (2001), The "Single Sales Factor" Formula for State Corporate Taxes: A Boon to Economic Development or a Costly Giveaway? Washington, DC: Center on Budget and Policy Priorities.

MCLURE, C.E., Jr. and W. HELLERSTEIN (2002), "Does Sales-Only Apportionment of Corporate Income Violate the Gatt?" NBER working paper No. 9060. Cambridge, Massachusetts: National Bureau of Economic Research.

MEYER, B.D. (1995), "Natural and Quasi-Experiments in Economics." *Journal of Business and Economic Statistics* 13(2): 151-161.

MOLNAR, L.A., D.R. GRIVES, J. EDELSTEIN, R. DEPIETRO, H. SHERMAN, D. ADKINS and L. TORNATZKY (with Y. BATTS, G. FULTON and S. HAYHOW) (1997), *Impact of Incubator Investments*. Athens, Ohio and Ann Arbor, Michigan: NBIA Publications and The University of Michigan.

MURNANE, R.J., S. NEWSTEAD and R.J. OLSEN (1985), "Comparing Public and Private Schools: The Puzzling Role of Selectivity Bias". *Journal of Business and Economic Statistics* 3:23-35.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (2002), "The Manufacturing Extension Partnership: Delivering Measurable Returns to Its Clients." Washington, DC: US Department of Commerce, Technology Administration.

NEXUS ASSOCIATES (1999), *A Record of Achievement: The Economic Impact of the Ben Franklin Partnership*. Belmont, Massachusetts: Nexus Associates, Inc.

OAKLAND, W.H. and W.A. TESTA (2000), "The Benefit Principle as a Preferred Approach to Taxing Business in the Midwest." *Economic Development Quarterly* 14(2): 154-164.

PAPKE, L.E. (1993), "What Do We Know About Enterprise Zones." In *Tax Policy and Economy*, Vol. 7, J.M. Poterba, ed. Cambridge, Massachusetts: MIT Press, pp. 37-72.

PETERS, A.H. and P.S. FISHER (2002), *State Enterprise Zone Programs: Have They Worked?* Kalamazoo, Michigan: W.E. Upjohn Institute for Employment Research.

PETERSON, P.E. (1981), *City Limits*. Chicago: University of Chicago Press.

POOLE, K.E., G.A. ERICKCEK, D.T. IANNONE, N.MCCREA and P. SALEM (1999), *Evaluating Business Development Incentives*. Washington, DC: National Association of State Development Agencies.

REED, W.R. and C.L. ROGERS (2000), "Measurement Error and Endogeneity in Studies of State Tax Policy and Economic Growth." Unpublished paper, Department of Economics, University of Oklahoma, Norman, Oklahoma.

ROGERS, C. and W.R. REED (forthcoming), "Quasi-Experimental Control Group Methods for Estimating Policy Impacts". *Regional Science and Urban Economics*.

ROHDE, S., J. CASH and K. AMMARMAN (1990), "Study of the Capital Access Program". Unpublished working paper, Michigan Strategic Fund, Lansing, Michigan.

SMITH, J.A.and P.E. TODD (2001), "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods". *The American Economic Review* 91(2): 112-118.

THOMAS, K.P. (2000), *Competing for Capital: Europe and North America in a Global Era*. Washington, DC: Georgetown University Press.

WASYLENKO, M. (1997), "Taxation and Economic Development: The State of the Economic Literature". *New England Economic Review* (March/April): 37B52.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 5

Four Directions to Improve Evaluation Practices in the European Union: A Commentary on Timothy Bartik's Paper "Evaluating the Impacts of Local Economic Development Policies on Local Economic Outcomes: What has been done and what is doable?"

by
Daniele Bondonio,
Department of Public Policy and Public Choice,
Università del Piemonte Orientale,
Alessandria, Italy

Introduction

The aim of this commentary is to place the techniques and methods reviewed and recommended in Timothy Bartik's paper "Evaluating the Impacts of Local Economic Development Policies on Local Economic Outcomes: What Has Been Done and What is Doable?" into the context of the European Union (EU) (looking in particular at the local/regional economic development programs co-funded by EU Structural Funds). Can we say, on this side of the Atlantic, that EU-sponsored local/regional economic development programs have been subject to rigorous evaluations, as Timothy Bartik can claim is the case for a number of US programs? Using the same methodological framework contained in Timothy Bartik's paper I argue that on this side of the Atlantic we could improve evaluation practices in four different directions:

1. being clearer on what rigorous evaluation is;
2. recording better data;
3. incorporating evaluation needs into policy design;
4. exploiting the heterogeneity of regional implementation designs across the EU.

It is appropriate that, as the focus of Timothy Bartik's paper is predominantly on evaluating business incentive programs, this commentary is largely based on my on-going experience in evaluating EU-sponsored investment incentives to small and medium enterprises (SMEs) in Objective 2 (Ob. 2) areas[1] and comparing these incentives with US Federal "Empowerment Zones". To my knowledge, however, current and past evaluations of incentive programs for SMEs in Ob. 2 areas are somehow representative of general evaluation practices adopted for other local economic development policies implemented in the EU.

Being clearer on what rigorous evaluation is

This section of the commentary needs the statement of a short premise. Prevailing evaluation terminology is slightly different between the US and the EU for local economic development programs (particularly as regards evaluations of programs co-funded by the EU Structural Funds). Terms largely used in the EU such as "*ex ante*", "*in itinire*" and "*ex post* evaluation" are not so common in the US evaluation literature. The reason for this is that what in the

EU is referred to as "*ex ante* evaluation" is often in the US referred to as "program feasibility assessment" and is strictly considered as a part of the set of broad programming activities needed to effectively design policy interventions. "*In itinere* evaluation" is very often labelled as "monitoring program activities". Readers should be clear that Timothy Bartik's paper is mostly concerned with what in the EU would be referred to as "*ex post* evaluation" (perhaps with some interview and focus group methods that in the EU would also be referred to as "*in itininere* evaluation"). In agreement with Timothy Bartik's terminology I will in this discussion adopt the term "evaluation" in place of "*ex post* evaluation".

As is clearly presented in Timothy Bartik's paper, three complementary approaches are most needed for rigorous evaluation of local economic development policies that provide business development incentives:

1. outcome impact evaluation on proximate dimensions of business activity;
2. estimating fiscal and employment benefits on the overall local economy through regional model (such as REMI[2] or Implan[3]);
3. surveys of clients and client focus groups to improve effective management of the program.

Outcome impact evaluation on proximate dimensions of business activity (such as employment or investment expenditures) [approach *a)*] is a core component of rigorous evaluation. It yields crucial evidence on the proximate effectiveness of the program by estimating how "what happened" differs from "what would have happened but for" the policies. Such impact estimates can then serve as a basis for estimating the fiscal and employment benefits of the program on the overall local economy through regional macro-econometric models that may allow an assessment of the cost-effectiveness of the intervention [(approach *b)*]. Surveys of clients and client focus groups, instead, can be run quite independently from both approaches *a)* and *b)* and constitute what would be referred to as "*in itinere* evaluation" in the prevailing terminology of EU Structural Fund evaluations.

In place of rigorous outcome impact evaluation, current evaluation practice for programs co-sponsored through the EU structural funds often attempt to measure only "what happened" in the target areas (or firms) instead of estimating differences between "what happened" and "what would have happened but for" the policies. If precisely measured, knowing "what happened" can be useful as it yields important information on the program activity that could help to effectively manage the program. However, as we are strongly reminded in Timothy Bartik's paper, it has to be clear that measures of "what happened" are not enough for rigorous evaluation.

Examples of such practice can be found in the thematic evaluation report on the Structural Fund impacts on SMEs commissioned by the European

Commission (Ernst and Young 1999) and two reports summarizing the European Commission's evaluation of the 1989-93 Ob. 2 programs (Malan 1998, Bachtler and Taylor 1999). In Ernst andYoung (1999) and Malan (1998), judgments on the impact of EU Ob. 2 programs on target areas are formed by comparing the economic growth of the Ob. 2 regions with growth in the group of non-Ob. 2 regions in the EU. However, such comparison would identify the actual impact of the Ob. 2 programs only if assisted and non-assisted areas would have performed in the same way without the intervention. Assigning all credit for any performance difference between assisted and non-assisted areas to the program does not constitute rigorous evaluation practice unless sound statistical/econometric methods (*e.g.* Bartik 1991, Bartik and Bigham 1995, Bondonio 2000, Bondonio 2002, Manski 1995, Moffit 1991, Smith 2000) are used to test and separate performance differences due to different pre-intervention characteristics between assisted and non-assisted regions.

In some evaluation reports commissioned by Italian regional administrations to evaluate the business incentive measures implemented in Italian Ob. 2 areas [*e.g.* Ecoter "Docup Ob. 2: Rapporto di valutazione finale" (Docup Ob. 2: Final Evaluation Report), prepared for the Piedmont Region, 1999] impact estimates of the program intervention are retrieved by summing the number of jobs that assisted entrepreneurs reported in their application packages would soon be created following the completion of the assisted investment.

Other evaluation practices have produced impact estimates of cash assistance programs for SMEs in Ob. 2 areas by counting the total number of jobs that interviewed entrepreneurs indicated as jobs that would have not be created absent the program assistance (*e.g.* Malan 1998 and Ernst and Young 1999). As argued by Timothy Bartik, such an approach might be biased by the tendency of assisted entrepreneurs to claim that cash benefits had a large impact, so as to keep the cash coming in the future.

It has to be noted, finally, that applying REMI, Inplan and other regional macroeconomic models to estimate fiscal and employment benefits on the local economy does not constitute rigorous evaluation if such models are estimated based on inputs that are unreliable measures of the program outcome impact on proximate dimensions of business activity. Lacking reliable evidence from rigorous outcome impact evaluation, impact estimates on the overall local economy would be inaccurate if they were based on the wrong inputs given to the macroeconomic regional models. For example, in the final evaluation report prepared by Ecoter for the Piedmont Region in 1999, the overall employment impact on the local economy of Ob. 2 areas is estimated by applying a set of multipliers from a macroeconomic regional model directly to the total figure for all approved investments in the area (a

procedure which implicitly assumes that no investments would have occurred in the area in the absence of the program intervention).

Recording better data

Almost every existing evaluation report of geographically-targeted business incentive programs sponsored by the EU structural funds stresses the need to obtain better data on program activity to improve the quality of evaluation. Improving the program monitoring systems is typically the suggested remedy in order to yield more precise and reliable data on the program beneficiaries and the amounts of the cash assistance and/or services offered to them (see for example Ernst and Young 1999).

However, appropriate statistical methods often exploit performance differences between assisted and non-assisted firms (or areas) and before and after the program intervention to retrieve reliable net outcome impact estimates of the program intervention (*e.g.* Bartik 1991, Bartik and Bigham 1995, Moffit 1991, Smith 2000). Thus, it has to be clear that having the best data solely on assisted businesses or target areas is not enough for rigorous evaluation. To properly use such statistical methods would instead require good data on both assisted and non-assisted businesses (or target and non-target areas). For the case of spatially targeted business incentive programs, moreover, rigorous evaluation would greatly benefit from data recorded at the plant level by national/European statistical systems that match employment information from employer records with socio-economic data on residents of small statistical geographic units.

In the EU, NUTS_3s[4] are the smallest current geographical units at which data from the official statistical systems are currently easily available with good reliability for comparisons across time. However, NUTS_3s are not small enough to allow the boundaries of important assisted areas (such as the Ob. 2 areas) to be precisely reconstructed.

In Greenbaum and Bondonio (2003), characteristics of assisted areas for the US Federal "Empowerment Zone" (EZ) programs and EU Ob. 2 areas were analyzed and compared. In the US, data that precisely matched all EZ boundaries could be retrieved by combining sets of census tracts (standard geographic units used by the US Census Bureau). However, for the EU no exact measure of Ob. 2 area characteristics could be easily retrieved. Only the availability of reliable data for small standardized geographic units (such as the NUTS_5s) would have allowed an acceptable reconstruction of Ob. 2 area characteristics. At present, however, NUTS 5 data are difficult to obtain and very unreliable for comparisons across time, as they are based on city administrative boundaries that frequently change over time.

As reliable panel data are important for rigorous analysis, evaluation practices would greatly improve in the EU as a result of building integrated statistical systems that yield easily accessible data sorted by small geographic units that remain stable over time (or for which changes are limited and easily traceable).

Integrated EU data systems should also include registries of assisted firms from all sources of public assistance sorted by EU nations and/or regions. Creating such registries is very much needed in order to ensure that assisted firms are compared to non-assisted firms and not to firms receiving public subsidies from sources other than EU sponsored programs. "*De Minimis*" rules that impose caps on the total amount of public assistance receivable by EU firms are slowly inducing administrations of individual EU countries to create registries of all subsidized firms. Integrating such regional registries in a unified easily accessible European archive would be of great help to enable more rigorous evaluation to be performed throughout the EU.

Incorporating evaluation needs into policy design

As reported in Timothy Bartik's paper, reliable data for rigorous evaluation can also be obtained by incorporating some evaluation needs into the policy design of local economic development programs. This option should be given proper consideration in the EU as the implementation of changes in European statistical systems may be drawn-out and expensive. One way to obtain reliable data for rigorous evaluation would be to designate assisted areas those boundaries exactly overlap the existing geographical units (or groups of geographical units) of EU statistical systems.

Implementing policies with experimental protocols would be another way to obtain data for rigorous evaluation. Very often, however, strong political reluctance to exclude needy areas and/or firms from public assistance undermines large-scale implementation of experimental protocols for local economic development programs. Nevertheless, some form of experiment could be acceptable and should be given proper consideration in the EU, in particular using the procedure that Timothy Bartik describes of random selection of firms for targeted marketing of the program. As suggested in the paper, such an experimental protocol can produce a significant difference in the program's usage rate between the group of treated firms (those receiving the marketing efforts) and the control group of firms not receiving the marketing efforts. Differences in the program's usage rate between the treated and the control group of eligible firms could be exploited to retrieve reliable impact estimates of the program intervention. As it does not cause any eligible firm to be arbitrarily excluded from program assistance, the implementation of such an experimental protocol might face minimal political resistance.

Exploiting heterogeneity of regional implementation designs across the EU

As stated in Timothy Bartik's paper, statistical analysis using control or comparison groups can give insights into why and how a program works, provided that sufficient variation in program designs is observed and accurately measured.

For EU-sponsored programs, plenty of variation in policy implementation designs exists across the different regions where the Ob. 2 programs are implemented. Such heterogeneity in policy design (in these and other EU-sponsored programs) should not be considered a threat to the validity of the analysis because it limits the comparability of evaluation results across the EU. Rather, it should be viewed as a great opportunity for testing the effectiveness of a variety of policy designs and differences in the generosity of the cash incentives and/or services offered to assisted firms.

To take advantage of such heterogeneity, appropriate statistical methods have to be implemented so that across-region variation in policy features is adequately operationalised and region or country-specific independent economic trends are controlled for and kept separate from the impact estimates of the region-specific policy features (*e.g.* Bondonio, 2002). Moreover, if plant-level data are available, it is important to note that the analysis can be implemented by separating the observed business outcomes (*e.g.* employment growth) into three components:

- change attributed to new firms attracted to assisted areas;
- change in incumbent firms;
- change from firms that cease to trade.

Sorting business outcomes in such a way can be very useful as it would allow investigation of whether certain policy features are more effective in attracting new firms rather than countering decline in existing production (or *vice versa*). Incentives that appear to be appropriate for attracting new firms could be recommended for target sites such as newly developed industrial parks, rather than sites where the main targets of the intervention are firms already operating in the assisted area.

Notes

1. Regions with declining industrial production eligible for EU-sponsored assistance.
2. Regional Economic Models, Inc. (REMI®), *www.remi.com*.
3. IMPLAN Group, *www.implan.com*.
4. NUTS stands for Nomenclature of Units for Territorial Statistics which is the five-tier hierarchical regional structure used to standardize the economic territories of

the EU. NUTS_3 areas are in the middle of the hierarchical structure and are formed by the set of geographic units composed of single second-tier sub-national jurisdictions (comparable in many aspects to US counties). NUTS_1 areas (which are the largest units of the hierarchical structure composed as groups of contiguous regions or states corresponding to the largest sub-national jurisdictions for each EU nation) and NUTS_5 areas (composed as the set of city or town jurisdictions of EU nations) complete the hierarchy.

References

BACHTLER, J. and TAYLOR S. (1999), "Objective 2: Experiences, Lessons, and Policy Implications", Final Report, European Policies Research Centre, *http://europa.eu.int/comm/regional_policy/sources/docgener/evaluation/pdf/finalrep_full.pdf*.

BARTIK, T.J. (1991), *Who Benefits from State and Local Economic Development Policies?* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

BARTIK, T.J., BINGHAM R. (1995), *Can Economic Development Programs be Evaluated*, W.E. Upjohn Institute for Employment Research, Kalamazoo, MI: Staff Working Paper 95-29.

BONDONIO, D. (2000), "Statistical Methods to Evaluate Geographically-Targeted Economic Development Programs", Heinz School Working Papers No. 2000-5, Carnegie Mellon University, *www.heinz.cmu.edu/wpapers*.

BONDONIO, D. (2002), "Evaluating Decentralized Policies: A Method to Compare the Performance of Economic Development Programmes across Different Regions or States", *Evaluation*, Vol. 8, No. 1, pp. 101-124, 2002.

GREENBAUM, R. and BONDONIO, D. (2003), "A Comparative Evaluation of Spatially Targeted Economic Revitalization Programs in the European Union and the United States", ICER-International Center for Economic Research Working Paper Series No. 3/03, January 2003.

ERNST and YOUNG (1999), "Thematic Evaluation of Structural Fund Impacts on SMEs", Synthesis Report, European Commission.

MALAN, J. (1998), "Translating Theory into Practice: Lessons from the Ex Post Evaluation of the 1989-93 Objective 2 Programmes", paper presented at the 1997 Seville Conference on Evaluation Practice in the Field of Structural Policies, *http://europa.eu.int/comm/regional_policy/sources/docconf/seville/sevil_en.htm*.

MANSKI, C.F. (1995), *Identification Problems in the Social Sciences*. Cambridge, MA and London, UK: Harvard University Press.

MOFFIT, R. (1991), "Program Evaluation with Nonexperimental Data", *Evaluation Review* 15, 3:291-314.

SMITH, J. (2000), "A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies", *Swiss Journal of Economics and Statistics* 136(3): 1-22.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 6

The Evaluation of Programs aimed at Local and Regional Development: Methodology and Twenty Years of Experience using REMI Policy Insight

by
Frederick Treyz, Ph.D.
and
George I. Treyz, Ph.D.

Foreword

Policy makers need to evaluate the total effect of local and regional programs in order to make informed decisions. Development proposals have economic, social, and demographic implications that go well beyond their direct effects. To understand these effects, analysts need to use a comprehensive economic forecasting and simulation model. Local and regional policy analysis models show the full effects of policy changes on the local economy, including socioeconomic consequences that may otherwise be unforeseen or unrecognized.

This paper describes the REMI Policy Insight model, the leading regional economic forecasting and policy analysis model. Over one hundred institutes, universities, government agencies and other organisations use custom-built REMI models specified to states, counties, cities, and other regions. These model users are located primarily in the US, but also include organisations using or planning to use models for regions in Belgium, France, Germany, Italy, the Netherlands, Spain, and the United Kingdom. The EU Commission has recently contracted REMI to develop REMI models for evaluation of structural fund investments.

Analysts use the REMI model to evaluate the economic effects of economic development programs, transportation infrastructure investments, environmental and energy regulations, and other policies that have an effect on the regional or local economy. REMI studies include evaluations of high-speed rail, new highways, business tax incentive programs, water resources issues, air pollution controls, electric utility deregulation, and hundreds of other applications. Often, users incorporate a REMI analysis into their process; for example, evaluating all potential business relocation proposals for a given state or city.

REMI provides a comprehensive modeling framework that shows total policy effects, even those that are not anticipated. For example, a policy to reduce air pollution may have cost consequences for businesses that will reduce competitiveness, but the policy may also have the non-pecuniary effect of making the area a more pleasant place to live. In this case, the loss of competitiveness will reduce output but the cleaner air will lead to inward migration that will increase the labor force. These increases will in turn increase labor productivity and reduce wages, thereby cutting costs and increasing competitiveness, which will increase economic activity. The net

effect of these policies can only be captured by a comprehensive model that includes economic migration, labor force, wage determination, land prices, and the location effect of competitiveness on employment and output in the area.

Another example of an unintended consequence that would require a model would be building a rail system with limited stops. This may increase competitiveness in selling services among areas with stations, but it may also decrease competitiveness in all services for each of the areas without stations. The net result of these changes has ramifications that may be unforeseen for the individual areas and industries and for the regional economy as a whole. Without an appropriate model, it may be impossible to predict the direction of economic activity changes in each of the areas in the model, either for employment in particular industries or in the economy as a whole, when both the transportation improvements and the cost of these improvements are combined.

Since the overall objective of economic development policy is to improve economic and social conditions, it is important to predict and evaluate the total effects of the policies in a systematic way. This will help in choosing the set of policies that will achieve the maximum benefit for the proposed expenditures. We have found that, to make this possible, a quantitative model must incorporate all of the key interactions in the economy and be based on solid economic theory. Such a model must also be designed to use relationships that are universal for market economies. It only requires the estimation of those parameters that cannot be determined by economic theory. These parameters can be estimated using data sets for a number of areas with similar behavioral characteristics. We have also learned that the software for such a model must be easy to use and enable the user to verify the reasonableness of the results.

The functioning of regional/local development analysis models used for policy/programme forecasting/simulation in the USA

Regional and local economic models are used in the US for a variety of policy analysis purposes, including evaluation of economic development proposals, transportation projects, environmental regulations, and energy programs. Regional models are also used for planning and programming purposes, particularly as they relate to infrastructure needs including new roads, airports, power plants, water facilities, and a broad range of other public and private services.

Practitioners use a number of methods for forecasting and policy analysis. Basic models include: input-output, computable general equilibrium, econometric, and new economic geography. Each method has its strengths and weaknesses. The REMI model, by integrating all of these methodologies,

provides a modeling framework that overcomes some of the weaknesses of using individual modeling methodologies alone.

Input-output models represent the way that the national, regional, and local economy operates through the interaction of various parts of the economy. Central to input-output analysis is the interaction of industry sectors in the economy. For example, to build an automobile, the motor vehicle manufacturer must use steel, tires, and other intermediate inputs in production. The use of labor and capital are also considered as inputs into production, and final demand for consumption, government spending, and exports are also represented in the input-output accounts. Typically, input-output models incorporate a high level of detail for large numbers of industries and twenty or thirty types of final demand.

Input-output models have several important limitations. First, these models assume that economic relationships are fixed and do not vary in response to cost changes, competition for resources, or supply constraints. Second, input-output models are static in nature and do not show dynamic processes over time. Third, they are often limited in their description of the economy; for example, input-output models do not provide for an interaction between population and economic changes.

Computable general equilibrium (CGE) models are usually based on well-formulated microeconomic foundations: firms maximize profits and households maximize utility. In contrast to input-output models, computable general equilibrium models are structured so that prices and wages respond to market conditions. Since the role of price and costs signals in the economy is fully specified, CGE models are particularly suitable for policy analysis regarding changes in taxes and production costs.

Computable general equilibrium models are used to evaluate changes in trade policy, environmental regulations, and taxes. For example, a CGE model may be used to show the economic effects of a state-level income tax increase. CGE models are not widely applied to economic development issues, such as the effect of a new firm location, since they usually do not have the industry detail or careful tracking of inter-industry relationships available in input-output models. Furthermore, CGE models, as commonly applied, often do not have an explicit time dimension. Therefore, policy analysis using such models takes the form of comparative static analysis, which can serve as a useful input into the policy making process but has significant practical limitations.

Econometric models rigorously employ statistical methods. Such models have an important basis in economic theory, and are more grounded in empirically supported relationships. Econometric models almost always have an important time dimension, and dominate other methods in quarterly and monthly short-term forecasting. Econometric models are sometimes used for

policy analysis purposes, but often are not appropriate for this purpose, as the economic structure may not be represented in sufficient detail for the types of exogenous shocks typical of policy applications.

Economic geography models have been developed in mostly a theoretical context in the last ten years. As with CGE models, this type of model is also based on clear microeconomic foundations but is typically much more stylized and not intended to represent an actual economy. Economic geography models are unique in that they are able to account for the endogenous formation of cities and other economic agglomeration.[1]

The REMI model brings together the methods of input-output, computable general equilibrium, econometric, and economic geography models in a comprehensive, consistent framework. The REMI model captures the inter-industry relationships embodied in input-output models. If all of the dynamic responses in REMI are turned off (as can be done through the "alternative model specification"), the model is specified as a static input-output model. The REMI model captures long term general equilibrium tendencies in labor and factor markets, as in computable general equilibrium models.

Dynamic responses in the REMI model are estimated using econometric techniques, typically using estimates based on panel data for various regions or states over a number of years. The model is also based on new economic geography theory, explicitly accounting for agglomeration economies in labor, input, and consumption markets.

Description of the structure of REMI[2]

REMI Policy Insight is a structural economic forecasting and policy analysis model. As mentioned above, it integrates input-output, computable general equilibrium, econometric, and economic geography methodologies. The model is dynamic, with forecasts and simulations generated on an annual basis in relation to behavioral responses to wage, price, and other economic factors.

The REMI model consists of thousands of simultaneous equations; however, its basic structure is relatively straightforward. The exact number of equations varies depending on the extent of industry, demographic, demand, and other detail in the model. The overall structure of the model can be summarized in five major blocks: 1) output and demand, 2) labor and capital demand, 3) population and labor force, 4) wages, prices and costs, and 5) market shares. The blocks and their key interactions are shown in Figures 6.1 and 6.2.

The output and demand block consists of output, demand, consumption, investment, government spending, exports, and imports, as well as feedback from output change due to the change in the productivity of intermediate inputs. The labor and capital demand block includes labor intensity and productivity as well as demand for labor and capital. Labor force participation

Figure 6.1. **REMI model linkages**
Excluding economic geography linkages

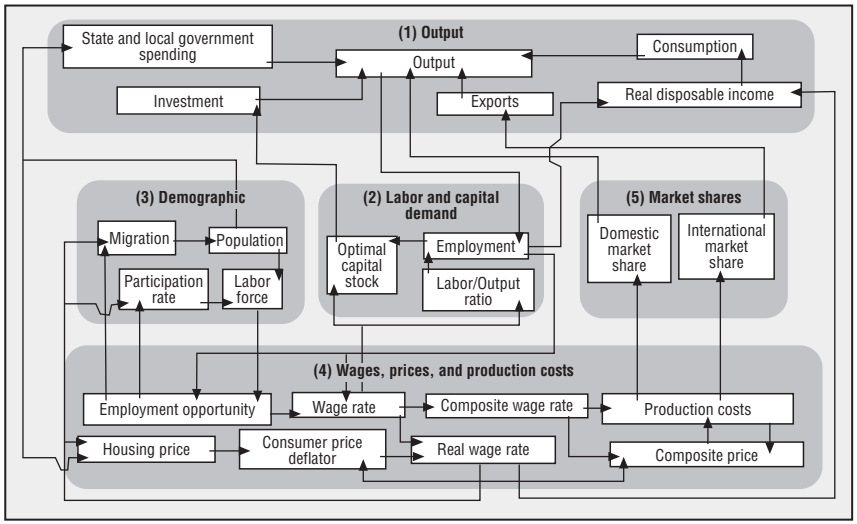
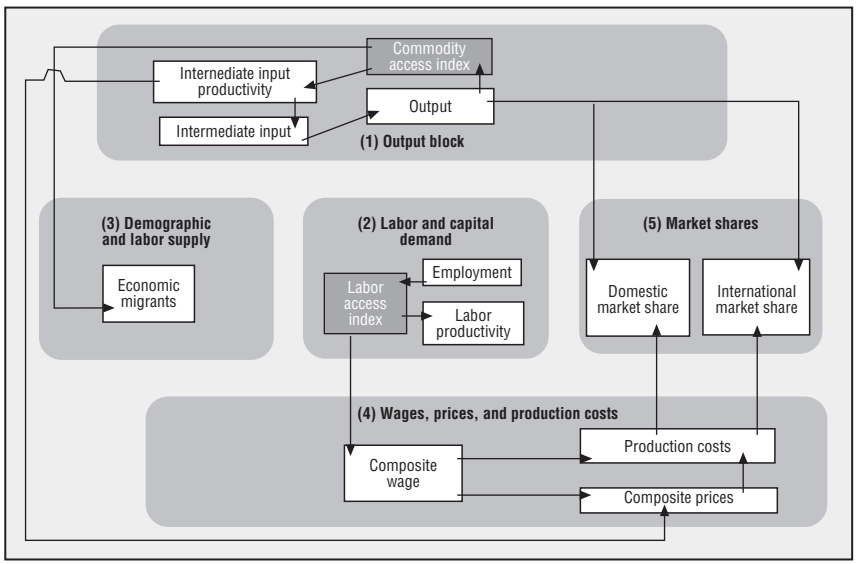


Figure 6.2. **Economic geography linkages**



EVALUATING LOCAL ECONOMIC AND EMPLOYMENT DEVELOPMENT – ISBN 92-64-01708-9 – © OECD 2004

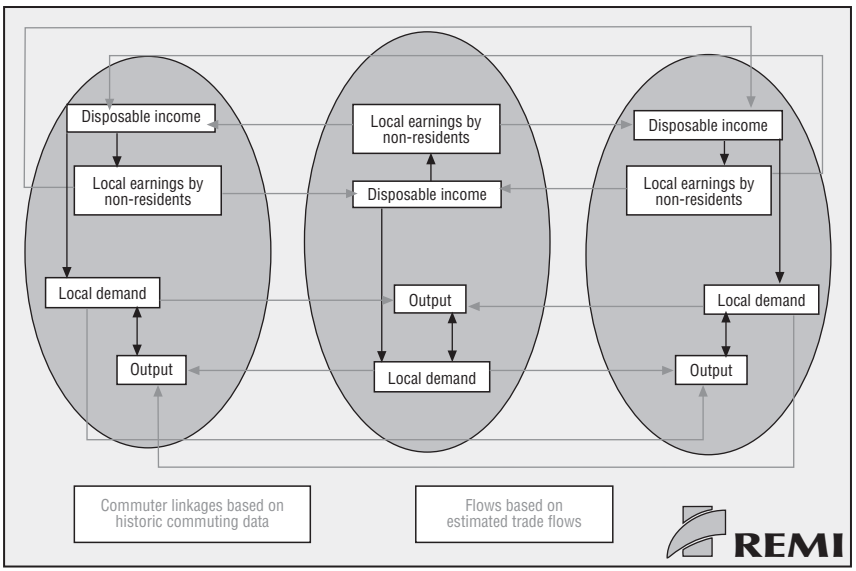
rate and migration equations are in the population and labor force block. The wages, prices and costs block includes composite prices, determinants of production costs, the consumption price deflator, housing prices, and the wage equations. The proportion of local, inter-regional and export markets captured by each region is included in the market shares block.

Models can be built as single region, multi-region, or multi-region national models. A region is defined broadly as a sub-national area, and could consist of a state, province, county or city, or any combination of sub-national areas. Within a large, multinational currency zone such as the European Union, models of a national economy can be built using the same economic framework employed in regional models.

Single region models consist of an individual region, called the home region. The rest of the nation is also represented in the model. However, since the home region is only a small part of the total nation, the changes in the region do not have an endogenous effect on the variables in the rest of the nation.

Multi-regional models have interactions among regions, such as trade and commuting flows. These interactions include trade flows from each region to each of the other regions. These flows are illustrated for a three-region model in Figure 6.3. There are also multi-regional linkages through delivered prices from each area to each other area as well as labor and commodity access among areas.

Figure 6.3. **Trade and communter flow linkages**



Multi-regional national models that encompass an entire currency union, such as the US or EU, may also include a central bank monetary response that constrains labor markets. Models that only encompass a relatively small portion of a currency union are not endogenously constrained by changes in exchange rates or monetary responses.

Block 1. Output and demand

This block includes output, demand, consumption, investment, government spending, import, product access and export concepts. Output for each industry in the home region is determined by industry demand in all regions in the nation, the home region's share of each market, and international exports from the region.

For each industry, demand is determined by the amount of output, consumption, investment and capital demand on that industry. Consumption depends on real disposable income per capita, relative prices, differential income elasticities and population.[3] Input productivity depends on access to inputs because the larger the choice set of inputs, the more likely that the input with the specific characteristics required for the job will be formed. In the capital stock adjustment process, investment occurs to fill the difference between optimal and actual capital stock for residential, non-residential, and equipment investment.[4] Government spending changes are determined by changes in the population.

Block 2. Labor and capital demand

The labor and capital demand block includes the determination of labor productivity, labor intensity and the optimal capital stocks. Industry-specific labor productivity depends on the availability of workers with differentiated skills for the occupations used in each industry. The occupational labor supply and commuting costs determine firms' access to a specialized labor force.

Labor intensity is determined by the cost of labor relative to the other factor inputs, capital and fuel. Demand for capital is driven by the optimal capital stock equation for both non-residential capital and equipment. Optimal capital stock for each industry depends on the relative cost of labor and capital, and the employment weighted by capital use for each industry. Employment in private industries is determined by the value added and employment per unit of value added in each industry.

Block 3. Population and labor force

The population and labor force block includes detailed demographic information about the region. Population data is given for age, gender and ethnic category, with birth and survival rates for each group. The size and

labor force participation rate of each group determines the labor supply. These participation rates respond to changes in employment relative to the potential labor force and to changes in the real, after tax, wage rate.[5] Migration includes retirement, military, international and economic migration.[6] Economic migration is determined by the relative real after tax wage rate, relative employment opportunity and consumer access to variety.

Block 4. Wages, prices and costs

This block includes delivered prices, production costs, equipment cost, the consumption deflator, consumer prices, the price of housing, and the wage equation. Economic geography concepts account for the productivity and price effects of access to specialized labor, goods and services.

These prices measure the price of the industry output, taking into account the access to production locations. This access is important due to the specialization of production that takes place within each industry, and because transportation and transaction costs of distance are significant. Composite prices for each industry are then calculated based on the production costs of supplying regions, the effective distance to these regions, and the index of access to the variety of output in the industry relative to the access by other users of the product.

The cost of production for each industry is determined by cost of labor, capital, fuel and intermediate inputs. Labor costs reflect a productivity adjustment to account for access to specialized labor, as well as underlying wage rates. Capital costs include costs of non-residential structures and equipment, while fuel costs incorporate electricity, natural gas and residual fuels.

The consumption deflator converts industry prices to prices for consumption commodities. For potential migrants, the consumer price is additionally calculated to include housing prices. Housing price changes from their initial level depend on changes in income and population density.

Wage changes are due to changes in labor demand and supply conditions and changes in the national wage rate. Changes in employment opportunities, relative to the labor force and occupational demand changes, determine wage rates by industry.

Block 5. Market shares

The market shares equations measure the proportion of local and export markets that are captured by each industry. These depend on relative production costs, the estimated price elasticity of demand, and effective distance between the home region and each of the other regions. The change in share of a specific area in any region depends on changes in its delivered price and the quantity it produces compared with the same factors for

competitors in that market. The share of local and external markets then drives the exports from and imports to the home economy.

Description of the data

REMI has two major uses for data. One is a use for time series data, when available, to estimate some of the coefficients of the equations. The other need for data is to calibrate the model to the year that will serve as the base historical year.

Estimating the coefficients

Due to the completeness of the structure in the REMI model, there is only a relatively small set of key equation coefficients to be estimated. This is possible because the model is based on the application of mainstream economic theory, including the recently developed new economic geography. The key assumptions underlying the model are that firms seek to maximize profits, households seek to maximize their well-being, and that workers as well as goods and services are not homogenous even if they are within the same occupation or sector of the economy. The equation structure of the model has been derived from these basic assumptions. In our 20+ years of experience, it has become evident that, due to the quality of regional data, the complexity of economics, and the information that firms and households consider, the larger the data set used the better the quality of the estimates. This is true even after the structure of the model has been designed in such a way that the number of coefficients and the data requirements are reduced to a minimum.

For the quantification at this basic level, the similarity of behavior for actors in different regions is more important to robust estimation than is the importance of slight differences from one region to another. In the United States this has led us to develop a database at the 53-industry level over 25+ years for 3 083 counties. This database obviously includes counties that run the gamut from small rural counties with only a few thousand people to populated urban counties. In the estimation process, we filter the data to insure accuracy for the industry in each county that is included in the industry-specific estimate.

In Europe, the time series data sets are shorter and often at a higher geographical level but provide us with a data set that makes it possible to make any necessary adjustments in our US coefficient estimates due to the EU specific conditions.

In the rest of this section we will discuss all of the necessary equation coefficients estimates in logical groups.

We start with two key coefficients used in the market share equations.

β_1 =The distance (measured as travel time) decay parameter in a gravity model for industry i.

σ_i =The price elasticity for industry i.

The β_1 coefficient determines how the home territories' share of the demand in the market of each other territory declines as the travel time to each of the other territories increases. The σ_i value shows how the home territories' share of each market (including its own market) will change in response to a change in the delivered cost from the home territory to each of the other territories. This is relative to the average delivered cost change in each territory based on its purchases from all of the other territories that it is purchasing from. The change in delivered cost from the home territory can be due to changes in its production cost or changes in the cost of delivery.

In order to estimate the β_1 (travel time decay parameter) dynamically, we need to have time series estimates for each of the following:

1. An approximation of the change in total output by industry for domestically produced goods and services in each territory for each year for each industry.
2. An approximation of the change in total demand for domestically produced goods and services in each territory for every industry and year. This is broken down into:
   a) Consumer demand.
   b) Investment demand (gross capital change).
   c) Government demand.
   d) Intermediate demand.

We can approximate the change in each of these concepts if we have changes in employment by industry, a national (or EU) input-output table, and a travel-time matrix from each territory to each other territory. We have obtained such a matrix commissioned by the EU for all NUTS II and NUTS III regions in the EU areas from Professor Wegener.[7] This is accomplished as follows:

1. An approximation of the change in output for domestic use in each territory for each year for each industry – use national outputs by industry from the national input-output table and national employment data (or use time series national output by industry and employment by industry if available), then apply the appropriate ratio to the local employment series.
2. A change in total demand by industry for each territory for each year.
   a) Consumption demand change – based on changes in the wage bill (or disposable income if available). Converted to industry demand using consumption vector in the national or EU input-output table.

b) Investment demand change – based on wage bill in construction (or changes in the rate of change in the Gross Capital Stock) converted to industry using the change in gross capital stock in the input-output table.

c) Government demand – changes in the total wages (or changes in population) convert demands from government from the private economy to industry demand using the appropriate national input-output table.

d) Change in intermediate demand for outputs of each industry based on output estimates in (1) above and the national (or EU) input-output table.

After this data is assembled, then the β_i (travel time decay parameters) can be found using a search algorithm across the β_i values to find the β_i values. This minimizes the rate of error for predicting the change in output in each territory, based on the changes in demand in each of the other territories for that industry's goods or services. For example, output for grocery stores should be closely related (*i.e.* a high share) to the changes in demand in the home area and contiguous areas and not related very closely (*i.e.* a low share) to the change in demand in distant areas. However, the output in the home territory of automobiles should be much more related to demand changes in all territories than to change in the home area demand (*i.e.* very low decay) due to travel time. This will yield a relatively equal share in all markets.

An alternative way to estimate the β_i values is possible if one or more sub national input-output tables are available for any year. This method provides another quantitative estimate for adjusting the β_i values in order to best explain the inflows and outflows from the local to the other regions based on one year of data. In the US, we examine the implications of each of the β_i values for the average travel time for the good or service in question relative to its cost. We also look at the proportion of the local area's demand served by the local output over a range of areas. We compared the proportions to those in similar industries. We estimate the travel time of typical customers for the particular good or service in question. After doing this analysis we use all of the information to modify some of the β_i values.

The estimate of the σ_i (elasticity of price response) uses the same time series data as set forth in the time-series approach for the β_i estimates set forth above. In this case, we use an algorithm to search over values of σ_i to find out what value of σ_i would improve the fit between output changes and demand changes based on the changes in shares that it would predict in the markets subsequent to the relative change in production costs (approximated by wage rate changes) in all of the areas. In other words, the elasticity of demand σ_i (the percentage change in the quantity demanded given the percentage change in relative delivered price) would be determined econometrically.

In the US, we pool similar industries into categories so that the estimated elasticity is identical for all of the industries in the broader category. We also examine the elasticities for reasonableness. For some of the industries, we change the filter criterion to filter out territories that do not have a substantial representation of the industry in question or have erratic time series, possibly due to data reporting and classification errors. Using these methods, we obtain statistically significant estimates for all industries that meet our own test for professional reasonableness.

Looking at the employment and wage data available and the national input-output tables that are available for EU countries, it appears that we will not have trouble building models for any of these regions.

We typically satisfy the data requirements of the EU models in the following four ways:

1. Our joint venture partners are provided with a list of our requirements and are requested to supply as much of the data as they can.
2. We go to the NewCronos database of Eurostat for data that is not supplied by our partners.
3. We extract data from the Internet and make our own inquiries from any other sources available. Even with these three sources, however, we often have missing data points and need to fill in these gaps using estimation procedures.
4. In cases where data for previous years are available, fitting a curve to the available data and extending that curve to the year of interest estimates the data for the history year of interest. When the data supplied is for a higher regional level than required, the available data is spread out using correlated variables that reflect regional variations. A similar spreading is done when the data available is at an aggregation of the required industries. In such estimation procedures, we normally have control totals at the national geographical level, at a sub-national level, and at some aggregation of industries.

The REMI Policy Insight Model is set up to use very detailed demographic data, which we have easy access to in the United States. For many European countries, the data is not available at the level of detail that we need. Therefore, there are instances when statistical methods are needed to fill in the missing pieces of data, and we have many means at our disposal to do this.

Usually, population by single age cohort is easily accessible, at least at the national level. If we have this, then we will use this data to spread out data at the regional level, if needed, by assuming at the regional level that each age group within five-year cohort will be in the same proportion to each other, as they are at the national level.

If we do not obtain local labor force, or don't have participation rate forecasts, we will use the participation rates predicted by Eurostat. Since Eurostat does not have the number of cohorts that we want, we must spread this data. What we have done in the past is to calculate participation rates from Eurostat's population and Eurostat's labor force. Then we use the same participation rates for each of the individual cohorts for ages 16 to 19 years old and likewise for each single age cohort over 75 years old. We will then use the 5-year age and gender groupings for the remaining cohorts.

In the future, we plan to use Eurostat's population and spread it using the population data we already have. We will then use Eurostat's labor force and spread it, using actual labor force data from a nearby country or the US labor force proportions, to spread 5 year age cohorts out by single age cohorts. We can then calculate the participation rates using these values. Next, we will use these Eurostat-calculated participation rates with the client-provided population data to calculate a labor force for the forecast. Eurostat only provides population and labor force forecasts for every five years, so we use a linear method to fill the participation rates for the intermediate years.

For migration data, we will use national data to spread the age cohorts at a regional level, or we can use US or another country's data, spread both to the national and the regional level if necessary.

In the absence of natality rate forecasts, we assume that natality rates remain constant over the forecast time period.

For employment by occupation data, we fill the percentages for one missing industry by using the values from the most similar industry. When we have no breakdown for manufacturing industries, we use the same value for all the manufacturing industries that we are given for the manufacturing industry as a whole.

Using the data that is available to estimate the β_i's and σ_i's in the country in question, we will examine these estimates in light of the US estimates and other EU estimates for consistency and reasonableness. Our criteria for examining the estimates are as follows: the statistical value of the estimates (*e.g.* their *t*-test values), the similarity of the estimates to those in the US and other EU countries, the consistency of the results from one industry to another, and the judgment of our contacts in the country in question. We use our professional judgment on whether to use the new estimates directly, to use other pooling combinations, or to find other ways to arrive at the best possible estimates. If no estimates for the particular county are possible due to a lack of data, we will use other EU estimates combined with US estimates for β_i and σ_i to make estimates for the country in question. Given the dominance of many industries by international firms, there is a high probability that EU elasticities of purchase responses would be fairly

consistent with their responses in the US; therefore, it may be reasonable to use the US estimates in some cases. Given our techniques of estimation, a time series estimate can be made with as little as two or three years of data. Thus, we will have a way to test our estimates in all of the countries under study to assure that our results are consistent with other countries.

The labor productivity equation uses labor cost σ_i's (estimates of labor heterogeneity), which are then used in the labor access index. These estimates were made based on the amount and cost of cross commuting of people in 1300 Traffic Analysis Zones (TAZ's) around Chicago. The data included the occupation and industry as well as the place of residence and the place of work of all those in the sample. We also had travel time and commuting cost estimates from every TAZ to every other TAZ. The estimates are quite robust and, since they represent industry hiring decisions, would be expected to be similar in any economy where there was a benefit in matching workers to their jobs for the benefit of both the firms and the workers.[8]

The economic migration speed of response is a key equation to estimate the European data. In order to estimate this response, we need time series on the real wage rates, preferably using a deflator that includes housing prices, the relative employment to labor force ratio, and the number of net internal migrants each year from region to region. These series, with the possible exception of the price index using housing prices, are needed to test for an expected difference in speeds of adjustment for the US and EU. In the absence of time series data, we could establish whether the speed of adjustment response should be modified from that in the US or other European countries by observing the migration flow data that is available.

The labor force equation has two coefficient sets. The first set is related to the labor force participation responses. This reflects changes in the regions' current employment relative both to a synthetic labor force using the national participation rates by cohorts and to the current population by age/gender cohorts. The other set is related to the relative real wage rate. In the absence of changes, the participation rate will follow that predicted in the nation for baseline. The coefficients are by 20 age cohorts for 2 genders. They could be collectively calibrated to each country if male and female participation rates are available for a number of years for the regions modeled. There may also be European studies of participation rate changes caused by real wage and unemployment rate changes. If these are available, they could be incorporated.

When unemployment rates are high, the general response of decreased participation rates captures (and the time of adjustment to these changes reflects) the long-term effects caused by continual high unemployment in some regions. In the aggregate, we would want to update this equation based on regional participation rate changes in Europe (or in the country in question)

based on changes in unemployment conditions and the real wage rate. If possible, we may also test the effects of changes in the real wage relative to subsidies available for those who are not working.

The housing (or land price) equation will be estimated with data from the area in question. The change in the price depends on changes in relative real disposable income and relative population density. The more difficult series to find is a housing price estimate. This often has to be estimated using a series compiled by the real estate industry.

Finally, the wage equation depends on the relative moving average of employment divided by the moving average of the labor force and the industry-weighted current occupational demand over its moving average. The data for this equation requires the industry and employment data mentioned in the first part of this section in addition to an industry/occupation staffing ratio matrix at the national level. Initial estimates show somewhat lower wage responses in the EU than in the US The wage response would need to be re-estimated for the countries in question, if possible.

Summary for estimates of coefficients

From available evidence, it appears that adequate data exists to estimate the necessary coefficients for building models for Europe, North America, and most other market countries. When particular coefficients cannot be estimated for a certain country, we will have quantitative evidence from other similar countries that could be used to modify US coefficients as required. We do see the need for further work to ferret out data that already exists but is not easy to obtain.

Model calibration

While REMI has reduced the number of coefficients that need to be estimated to a bare minimum, the calibration to a designated last history year is still an additional task. It involves using all of the information that we have mentioned above but only for one year. It also involves making a national forecast for the country or monetary union in which the region is located. We can do this forecast ourselves, but are be willing to align it with an official forecast if necessary.

In making our national forecast, we first develop a national labor force forecast by applying participation rates to cohorts that are consistent with the Eurostat's projected participation rates. The combination of our projected productivity growth by industry and the size of our labor force sets a limit of output to maintain a fixed employment/labor force ratio for the baseline national forecast. An upward trend in the employment/labor force forecast could be built in if it were desirable to assume that current unemployment rates exceed their likely long-term average rates.

We have produced models of European countries with 24, 26 and 30 industries. In each case, we go to the smallest size region possible (usually NUTS II or NUTS III) to prepare a database that encompasses the entire country. This allows us to properly estimate the trade flows at that geographical level. Then we aggregate these flows for a limited number of equations and calculate the value of a concept, such as effective distance (travel time), in such a way that internal flows and the flows from one major region to another will be consistent with the flows among the smaller regions.

The demographic information necessary for forecast includes detail that we require for only the region being modeled. This is due to our need to have a structured demographic model with all the standard demographic processes.

Cost of model construction/use

As implied from the discussion above, REMI Policy Insight® is different and custom-built for each area we build a model for. Currently, our European models range is from 24-30 industrial sectors. These include a model of the Grampian region in the UK, a two-area model of France and the rest of the EU, a three-area model of the Netherlands, a two-area model of Rhine-Westphalia, and a single-area model in Tuscany. The latter three models involved ECORYS, RWI (Rhine-Westphalia Institute for Economic Research, Essen), and IRPET (Istituto Regionale Per La Programmazione Economica Toscana), respectively. REMI has recently been chosen by the European Commission to develop our model methodology for assessing the regional impact of structural funds. We will be building a model for Southern Italy and a model of four contiguous regions in Spain as examples for the EU Commission.

REMI Policy Insight is a licensed product customized to the region or regions for which it is intended. The price is 46 000 euros for a single-area model of any size region. The price for a two-area model is 53 000 euros. Each incremental region up to ten regions is 7 000 euros to yield a price of 112 000 euros for a ten-area model. For each of the next 10 areas the increment is 3 500 euros. After the first year, a 30% maintenance cost provides a new model each year and continued unlimited telephone and e-mail help in using the model and interpreting the results.

The use of program evaluations or *ex ante* projected direct effects as inputs for the REMI model

REMI Policy Insight® is a macroeconomic forecasting and policy analysis instrument. It includes a baseline forecast and provides policy variables such as employment, productivity, taxes, and production costs that can be changed by the user. Analysts may obtain the program evaluation policy variable inputs from the outputs of limited program evaluations. They may obtain similar

information from proposed direct interventions for projects under consideration.

For example, the direct outputs from traditional cost-benefit analysis of transportation investments provide the policy variable inputs needed for transportation studies. These direct variables include such factors as construction expenditures, operation expenditures, and travel time savings for businesses and consumers. These variables can be used as inputs into the REMI model to show the total macroeconomic effects of transportation infrastructure improvements.

The REMI model has also been linked to specialized microeconomic models. The Energy 2020 model,[9] which provides a high level of detailed energy-related information, allows users to simulate the effects of different types of electric generation plants and other energy policies. The outputs of Energy 2020, including utility construction spending as well as commercial, industrial, and consumer fuel price changes, are then used as inputs in the REMI model. The REMI model in turn supplies the predicted change in outputs that are required by the Energy 2020 model.

Overview of policy analysis areas

The REMI model is a comprehensive forecasting and policy analysis model. It includes thousands of economic and demographic policy variables as well as a complete description of the regional economy. Analysts are therefore able to use the model to evaluate a broad range of policy options. These include economic development programs and incentives, transportation investments, environmental and energy regulations, and other policies that have an effect on the economy. This section describes applications in several important areas, with a few illustrative examples that draw on the thousands of policy analyses that have been conducted using the REMI model.

Economic development

A broad range of government transfer programs, infrastructure investments, and business incentive programs are designed for the purpose of advancing the economic development of local and regional areas. Economic development as a governmental objective often targets the attraction or retention of specific firms. Also, infrastructure improvements such as an airport expansion often have an important economic development component.

The REMI model has been used for a broad range of economic development projects. These include issues ranging from the effects of the horse racing and breeding industry in Minnesota[10] to the impact of a

convention center and hotel in Kansas City[11] to the economic impact of casino gaming proposals in various states.[12] Analysts have evaluated the economic effects of automobile assembly plant locations in Kentucky,[13] Michigan,[14] Illinois, North Carolina, and Texas. Another theme is the economic effects of sports stadiums; studies include a football complex in Hartford, Connecticut and a new baseball park in Boston, Massachusetts.[15]

Policy makers often evaluate the economic effects of potential losses of key industries or employers. For example, REMI users have evaluated the regional economic effects of military base closures,[16] declines in tobacco sales,[17] lost coal sales,[18] and plant closures.[19] Researchers evaluated the economic effects of land use and growth controls in California. REMI users have also evaluated economic losses that would occur due to actual or potential natural disasters such as floods and hurricanes.[20]

To evaluate a new firm, the analyst typically needs to consider the construction phase, operational requirements, and government tax or spending incentives that may have been required to attract the firm. The construction of a facility occurs for a relatively brief time period, and usually generates many temporary jobs. Analysts can enter construction in the REMI model using either a general construction policy variable or policy variables relating to specific types of construction such as new industrial buildings, new office buildings, or new commercial buildings.

The user enters the direct effects of the operations phase of a new firm by using either employment or output policy variables. In using the employment policy variable, the analyst enters the total number of workers that are directly employed by the new firm. The employment policy variable assumes a given level of labor productivity, thus the model adds the direct output associated with these employees. Similarly, when the user enters output values, the model calculates the additional direct employment that is needed to produce this output. The user enters the direct employment or output values before running the model to show the total economic effects of the new firm.

As part of a targeted economic development policy, government agencies often offer specific infrastructure investments or tax incentives, or both, as a means of attracting businesses. For example, a state may provide a new exit ramp off of an interstate highway as part of a package to attract a manufacturing facility. More commonly, states and cities may offer corporate profits, property, and other type of tax breaks in order to bring a facility to a state.

To appropriately capture the economic effects of firm attraction, the analyst needs to explicitly consider the effects of government investments and/or tax incentives. For a government infrastructure investment, the analyst needs to add the infrastructure spending, and elsewhere reduce government spending or increase taxes to pay for the infrastructure

investment. For a tax incentive, the analyst needs to balance the government budget by increasing taxes or reducing government spending elsewhere in order to offset the tax cut.

In some cases, economic development officials may claim that a business tax incentive that takes the form of a tax exemption does not necessarily imply that taxes must go up or government spending be reduced elsewhere. These officials argue that, since a tax exemption involves no government transfer, it does not require any further government tax or spending changes. For most tax exemptions, this argument is fallacious. In general, the provision of government services is closely linked to economic activity. Even businesses that have few government services requirements often employ workers whom, as residents, require a high level of educational, fire and safety, and other public services. Thus, an increase in economic activity requires an increase in taxation in order to support the need for a higher level of government services. Without further information, the analyst should assume that the additional revenues generated by the firm are needed, at least in part, to pay for additional government service requirements.

REMI model users can implement government infrastructure changes using policy variables for detailed investments in highways, water and air facilities, and other detailed expenditure categories. Users can represent government tax changes with policy variables for personal income taxes, property taxes, sales taxes (for 13 consumption commodities), and a number of business taxes.

Some economic development initiatives have special aspects that can be evaluated using the REMI model. A seasonal or one-time major event such as a festival or sporting event is often pursued as a means of creating jobs and economic activity for a city or region. The University of Connecticut evaluated the economic effects of the proposed 1995 Special Olympics for New Haven, Connecticut.[21] Since much of the construction in preparation for the event and employment during the event is temporary, the greatest economic effects are transitory. Long-term effects result from the additional facilities that are available to the region, as well as possible business-location effects caused by the publicity generated by the event.

Universities and other institutions are often regarded as catalysts for the economic development of a region, not only due to the immediate employment and spending of the university, but also through increasing the productivity of the labor force and by acting as an incubator for new technologies and enterprises. Nexus Associates evaluated the effects of Tufts University School of Veterinary Medicine on the Boston and Massachusetts economy. In particular, the study traced the economic effects of spin-off businesses created from University research.[22]

Some economic activities that could be considered unattractive for a region may be valued due to their economic development benefits. Mining operations, electrical generation plants, paper mills, and other environmentally damaging activities may nevertheless be valued due to their job-creation potential. Prisons, often viewed as a disamenity due to safety concerns, are eagerly sought as an economic development engine for rural, economically depressed regions. For example, the Kentucky Governor's Office for Economic Analysis evaluated the economic effects of prison location in various locations in the mountainous, rural area of eastern Kentucky. The communities in this region actively pursued the prison location in their respective areas due to the stable employment that is provided by prisons but is otherwise unavailable in Appalachian Kentucky.[23]

Transportation

Analysts use the REMI model in order to evaluate the total economic effects of transportation projects. This analysis is often used as a means to go beyond the traditional direct cost and benefit estimates to show the wider repercussions for the region. Analysts have used REMI for hundreds of transportation studies involving highways, rail lines, ports, and airports.

Wilbur Smith Associates used the REMI model to conduct an *ex post* evaluation of the Appalachian Development Highway system. This system was constructed over a more than thirty-year period in order to increase accessibility of remote communities in the Appalachian mountains of the US The analysts used outputs from detailed transportation models as inputs into the REMI model. The direct changes, implemented as REMI policy variables, included competitive cost changes, roadside expenditures, and travel and tourism expenditures. The total economic effects of these direct changes included year-by-year estimates of increases in gross regional product, employment and wages.[24]

A study of a high-speed rail link between San Diego and San Francisco, California, used the REMI model to show the feasibility of the project based on benefit/cost and net present value measures.[25] The researchers considered a number of direct effects of the policy changes, which they modeled using appropriate REMI policy variables. Detailed railroad-specific construction and operation policy variables were used to represent the direct spending effects. Since the rail line would be used in place of other forms of transportation, thus reducing congestion, the analysts lowered transportation costs for highways, conventional rail, and aviation. The price of gas was raised in order to represent tax increases for gasoline to help finance the high-speed rail line. The analysts reduced housing costs in the coastal region of California and increased housing costs in the central region as a way of representing housing price shifts that would occur due to increased commuter accessibility to the

interior of the state. The analysts also used detailed consumer and government policy variables to represent shifts in spending patterns.

Energy and the environment

Energy and environmental economists use the REMI model to evaluate the economic implications of a broad range of policies. The US Environmental Protection Agency is using the model to show the potential economic benefits of state-sponsored programs to reduce the production of greenhouse gasses.[26] REMI medium- and long-term forecasts are used to drive models that estimate emissions, which then help to identify the needs for air pollution controls to reduce these emissions. Other environmental applications of REMI include the economic effects of reducing acid rain,[27] low-emission vehicle programs,[28] lake remediation,[29] and the restoration of the Florida everglades.[30]

Energy applications of REMI include the effects of an early shutdown of a nuclear power plant,[31] an evaluation of the Vermont comprehensive energy plan,[32] and the effects of demand-side management energy efficiency programs.[33] In an evaluation of electric market deregulation in Wyoming, analysts showed the effect of electric rate increases on the state. The analysts predicted that electric rates would increase as utilities would be able to market their relatively low-cost electricity at a higher market rate that would prevail with deregulation. The cost changes were implemented separately for commercial, industrial, and consumer cost changes in the REMI model. The effect of electric rate increases for businesses is to reduce business competitiveness and cause a substitution away from electricity towards higher use of labor and capital. Consumer electric rate increase result in lower real incomes, causing a small relative out-migration effect. The consumer price shifts also cause a substitution away from electricity towards other consumption categories.[34]

Applications of the REMI model: examples

Policy analysts use the REMI model as an integral aspect of planning and evaluation. Often, the use of the model fulfills an organisational mandate for analysis. For example, the South Coast Air Quality Management District in southern California uses the REMI model to meet a requirement for the evaluation of the socio-economic effects of all proposed air quality regulations. In another case, the Texas State Comptroller campaigned on the platform of implementing dynamic tax analysis for the state. The Comptroller selected the REMI model as a means of fulfilling this campaign pledge.

Several state-level economic development departments use the REMI model to measure the effects of various business attraction incentive programs. These departments often have the authority to offer tax breaks as a

means to encourage firms to locate in the state. However, these incentives are limited, so that firms must reach a threshold level of job creation in order to be eligible for such credits.

Illinois, Michigan, Kentucky, and Missouri are several of the states that have the REMI analysis as a standard part of their incentive package evaluation. Michigan, for instance, gathers information on the direct output and/or employment for the firm, and other firm-specific details such as wage rates and investment plans. The state evaluates the effectiveness of the incentives based on changes in total employment, wages, gross domestic product, fiscal effects for state and local government, and per capita real income. Based on this information, Michigan will either approve or deny the tax incentive.

Case study: the Mazda and the Mitsubishi automobile assembly plants[35]

Researchers at the University of Michigan pioneered the use of the REMI model for the evaluation of state-level tax incentive granting processes. The university, as well as state agencies, has developed and refined this type of procedure in order to rationalize the tax-incentive application process and provide maximum benefit to the public. Two early cases in Michigan were particularly important in providing a foundation for application methods, and serve as useful case studies to illustrate the role of REMI analysis in the policy making process.

In the mid-1980's, Mazda planned to build a new automobile assembly plant. The firm approached several states for assistance. As part of this process, researchers at the University of Michigan evaluated both the locational issues regarding the plant location in Michigan and the economic impacts of the firm location in Michigan. This discussion focuses on the economic impact methodology, which was central in determining the state incentive package.

The Mazda study assumed that the automobile assembly plant would employ 2 500 workers. This estimate was based on comparable plants elsewhere in the US The analysis also included investment expenditures to modify the existing Flat Rock plant for automobile assembly; this investment expenditure was assumed to take place over a three-year time period.

The results of the analysis show a creation of a total of 2 357 jobs in the first year of construction, increasing to 13 422 jobs during the first year of operation. The total job creation was predicted to begin to stabilize as an increase of 12 684 and 12 422 jobs in the subsequent two years of operation, respectively. The study also provided a number of sensitivity tests, such as an evaluation of the wage effect on increases in total employment.

The state provided a tax subsidy as part of an effort to attract the Mazda facility. In a related case, the University of Michigan researchers evaluated the economic impacts of the potential location of a Mitsubishi plant in the state. Mitsubishi asked for incentives from Michigan, Illinois and several other states as part of choosing the plant location. In this case, based on the economic impacts, the Michigan researchers suggested that the tax incentive package asked for by Mitsubishi was too great. As a result of their recommendation, the state withdrew from the competitive bidding for the plant. As a result, the state saved millions of dollars from tax incentives that might otherwise have been given away without justifiable economic benefits. Mitsubishi ultimately located the plant in Illinois.

Sample simulations

To show the performance of REMI Policy Insight, we have prepared simple sample policy simulations for several examples. In these examples, we have only made changes in one or two policy variables for illustrative purposes. Almost all realistic policy simulations require the user to change a number of policy variables to reflect details of the proposed policy.

International immigration

This simulation was constructed by increasing the level of international migration for all age groups. As the level of international migration increased due to the policy assumption, there was net out migration of some of the existing labor force in response to the increase in labor supply (and subsequent decline in relative employment opportunities and relative wage rate caused by the additional international migration).

The additional labor force provided by the increase in international migration is shown to decrease the cost of production by 0.091% by 2010 as the labor availability causes relative wages to be somewhat lower than they would have been otherwise. Figures 6.4*a* and 6.4*b* show the policy variable inputs used, in this case, by 133 thousand per year. Figure 6.5 shows the population increase for which there are three components: the increase in international migration caused by the user assumption, a partial offsetting decline in economic migration in response to the labor force competition from international migration, and the cumulative net population change over the simulation horizon. As shown in Figure 6.6, the output of New York City in this simulation increases compared to the baseline. This is driven by an increase in exports, which occurs as businesses become more competitive due to a larger labor supply and the resulting relatively lower wage rates. The unemployment table, Figure 6.7, shows that the labor force grows faster than employment, resulting in an increase in the unemployment rate. The detailed participation

Figure 6.4*a*. **Policy variable inputs used**

Figure 6.4*b*. **Policy variable inputs used**

Figure 6.5. **Population determinants (differences)**

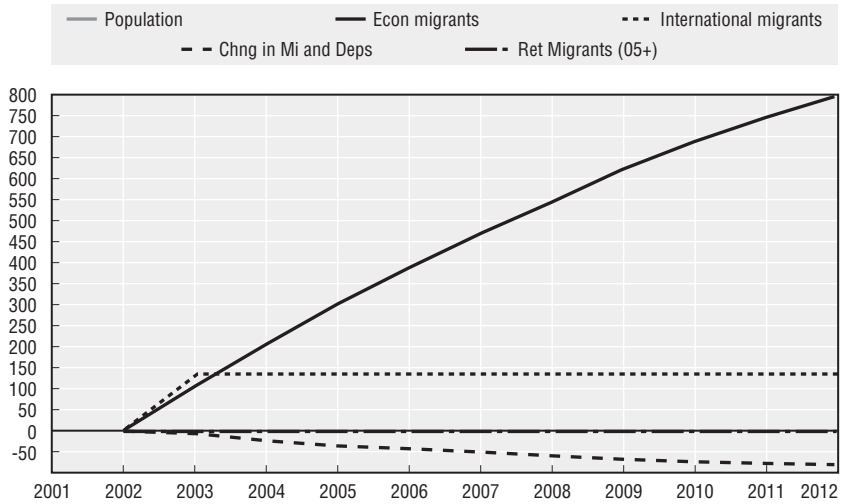


Figure 6.6. **Output components by demand source (per cent change)**

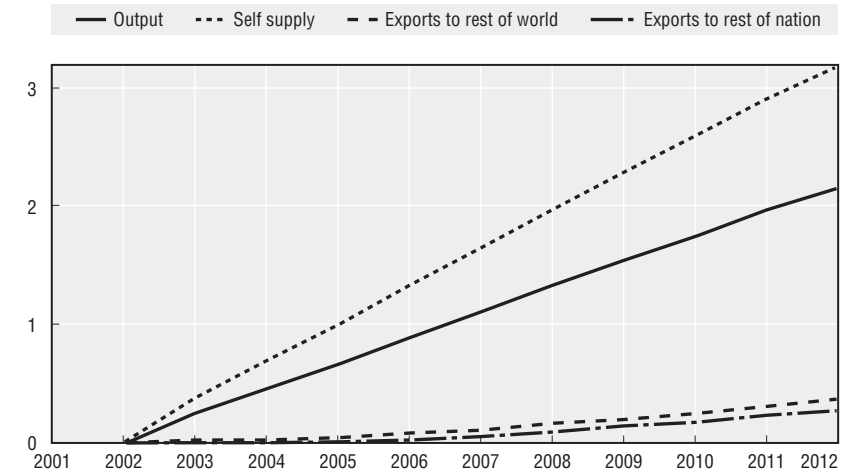
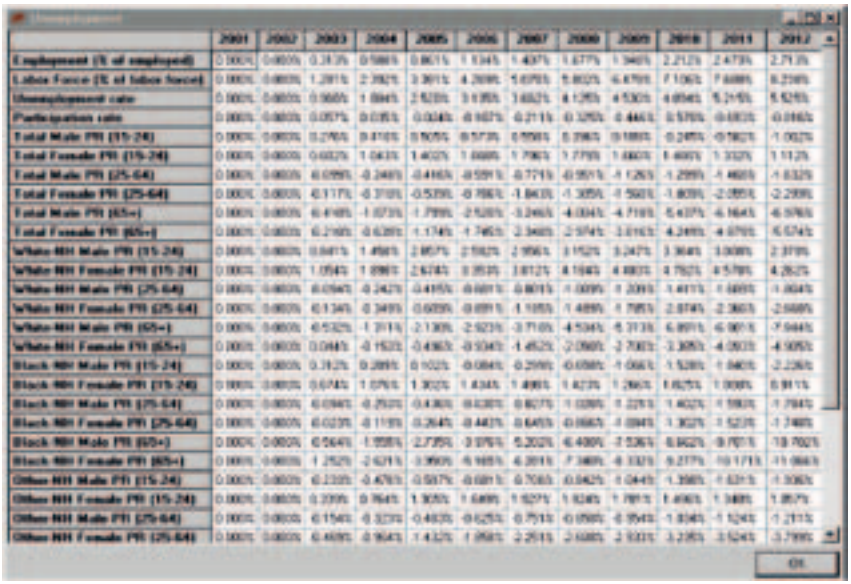


Figure 6.7. **Unemployment table**



rates show changes for different groups. These changes are caused by real wage and unemployment rate changes as well as changes in age composition within the groups due to the age structure of international migrants.

A permanent one per cent increase in total factor productivity in 2003 for Greater Rotterdam

A simulation using a three-region model of the Netherlands shows the economic response to an increase in total factor productivity (TFP). TFP was assumed to increase for Greater Rotterdam as a permanent 1% increase in TFP starting in 2003.

Figure 6.8 shows the response in output by demand source for the Greater Rotterdam region. As TFP increases and enhances the competitiveness of Rotterdam industries, their sales to the rest of the Netherlands and international markets increases. In response to the once and for all change in 2003, the increase in self supply is greater than 2% in the long-term, whereas exports to the rest of world and multiregions increase by just over 1%.

Figure 6.9 shows employment by demand source for Greater Rotterdam. In the short-term, there was a drop in employment as the greater productivity requires fewer workers for the same level of output. However, the increase in competitiveness caused by TFP increases results in a positive increase in total

Figure 6.8. **Output components by demand source (per cent change)**

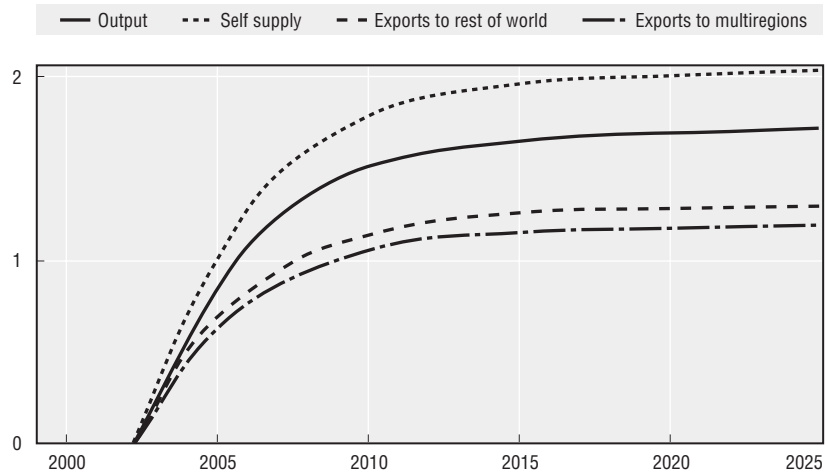
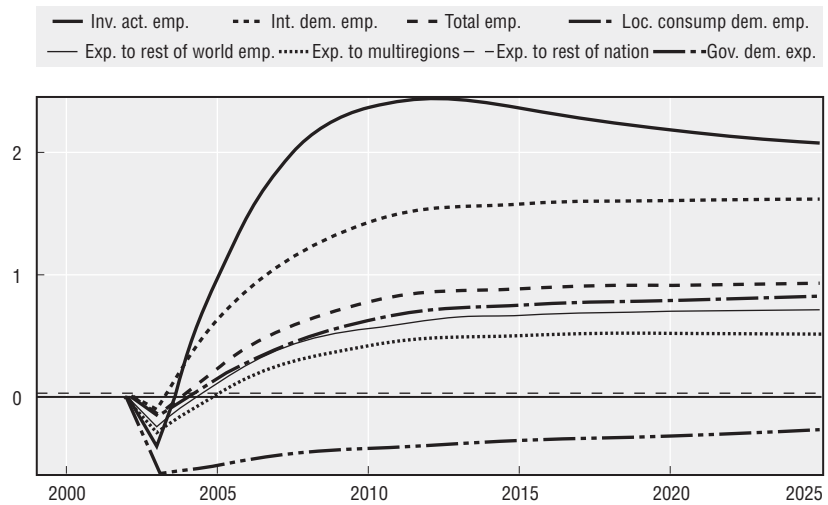
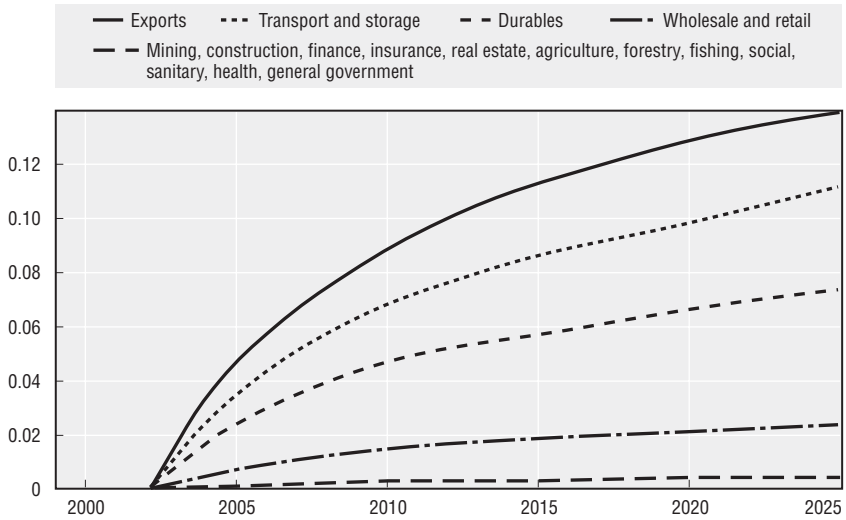


Figure 6.9. **Employment components by demand source (per cent change)**



employment after 2 years. This is driven by the increase in employment due to increased exports and also to the capital stock adjustment process. In this process, investment activities and investment-related employment increase rapidly as the economy expands. The increase in investment tapers off somewhat as the capital stock is built up to a new level.

Figure 6.10. **Exports to rest of world (differences)**



Trade flows between the regions in the model are shown in Figure 6.11. For a specialized industry constructed of wood, furniture, etc., this table shows that Greater Rotterdam provides more of its needs for this industry while it imports less from the rest of the world. There is also an increase in trade among all Netherlands regions due to the expansion of the national economy as centered around the Greater Rotterdam economy.

Figure 6.12 summarizes the changes for Greater Rotterdam. Increased productivity in Rotterdam industries leads to a decrease in the cost of production, delivered price, and the price index. Enhanced competitiveness leads to an increase in exports, which drives the increases in employment, gross regional product, and personal income.

Figure 6.11. **Trade flow matrix (differences)**

2003 | Wood, wood prod, furn, jewelry, games, toys | Type: Differences

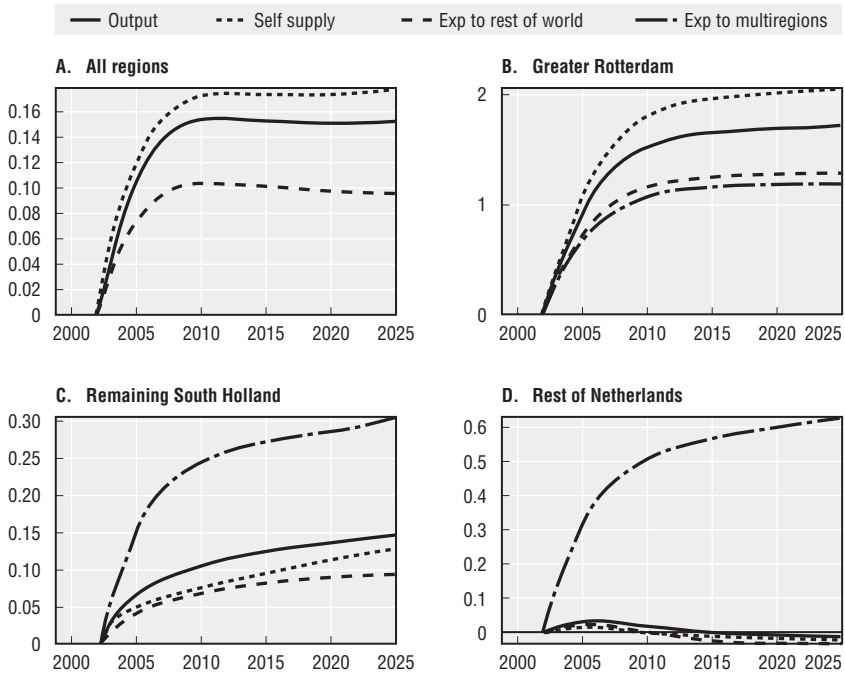
	Greater Rd	Remaining	Rest of Ne	Rest of Na	Rest of W	Output
Greater Rotterdam	0.0020681	0.00027004	0.00051905	0	0.00072996	0.0035871
Remaining South Holland	0.00039314	0.00026566	0.00014043	0	5.6744E-5	0.00086212
Rest of Netherlands	0.0012852	0.00028682	0.0018396	0	0.00016403	0.0035763
Rest of Nation	0	0	0	X	X	X
Rest of World	-0.00034183	0.0002867	0.00093746	X	X	X
Demand	0.0034106	0.0011094	0.0034361	X	X	X

Figure 6.12. **Summary table (per cent change)**

Variable	2003	2004	2007	2010	2020
Employment (Thous)	+0.015%	+0.020%	+0.020%	+0.010%	-0.013%
GRP (1,000M 98eur)	+0.014%	+0.019%	+0.016%	+0.003%	-0.028%
Pers Inc (1,000M Nom eur)	+0.007%	+0.010%	+0.011%	+0.006%	-0.008%
PCE-Price Index (NL 1998=100)	-0.011%	-0.009%	-0.005%	-0.001%	+0.001%
Real Disp Pers Inc (1,000M 98eur)	+0.017%	+0.018%	+0.014%	+0.007%	-0.009%
Population (Thous)	+0.000%	+0.000%	+0.001%	+0.001%	-0.001%
Total Migrants	+0.053%	+0.049%	+0.030%	-0.005%	-0.064%
Labor Force	+0.002%	+0.003%	+0.006%	+0.005%	-0.004%
Demand (1,000M 98eur)	+0.018%	+0.023%	+0.025%	+0.016%	-0.007%
Output (1,000M 98eur)	+0.017%	+0.024%	+0.028%	+0.017%	-0.010%
Delivered Price	-0.013%	-0.011%	-0.006%	-0.002%	+0.001%
Rel Cost of Production	-0.007%	-0.005%	+0.000%	+0.004%	+0.007%
Labor Intensity	0.000%	0.000%	0.000%	0.000%	0.000%
Labor Access Index	0.000%	0.000%	+0.001%	+0.002%	+0.004%
Indust Mix Index	0.000%	0.000%	0.000%	0.000%	0.000%
Reg Pur Coeff (SS over Dem)	-0.002%	-0.004%	-0.010%	-0.016%	-0.027%
Imports (1,000M 98eur)	+0.022%	+0.031%	+0.045%	+0.047%	+0.042%
Self Supply (1,000M 98eur)	+0.016%	+0.020%	+0.015%	0.000%	-0.034%
Exports to Multiregions (1,000M 98eur)	+0.103%	+0.210%	+0.414%	+0.507%	+0.601%
Exports to Rest of Nation (1,000M 98eur)	N/A	N/A	N/A	N/A	N/A
Exp to Rest of World (1,000M 98eur)	+0.008%	+0.012%	+0.014%	+0.005%	-0.020%
Wage Rate (Thous Nomeur)	+0.001%	+0.002%	+0.003%	+0.002%	-0.004%

Figures 6.13*a*–6.13*d* show the per cent change in output components by demand source for all regions in the Netherlands, the Greater Rotterdam region, Remaining South Holland, and the Rest of the Netherlands. Although major expansion in market shares to the rest of the world occurs in exports, the increase in self supply is a more significant increase due to the expansion of the local economy. This is shown for the changes in self supply and exports to the rest of the world. Figure 6.13*c* shows an increase in exports to other parts of the Netherlands from Remaining South Holland. Figure 6.13*d* shows the changes from the Rest of the Netherlands. Both Figures 6.13*c* and 6.13*d* show that, although the increase in productivity occurred in Greater Rotterdam, the expansion of the Greater Rotterdam economy leads to some of the import needs of Greater Rotterdam to be met by increased multiregion exports from other regions in the Netherlands model.

Figure 6.13. **Output components by demand source (per cent change)**



A reduction in travel time due to an improvement in transportation infrastructure

Finally, we present a simplified simulation with the European version of the REMI Policy Insight Demonstration model. This is a single area model of a small area in the UK. We do this to illustrate how REMI Policy Insight makes it possible for us in testing and evaluating the models to show all of the details that document and explain the results.

In this case we assume that the intervention is a transportation infrastructure investment that will decrease the internal travel time in the Demo region by 10% for 50% of the vehicle miles traveled. We simplify the analysis by entering an immediate reduction in transport and accessibility costs by 5% in the initial and all subsequent years. The information for making this type of estimate would come directly from a cost-benefit analysis in the proposal for the investment and/or from analysis of the location and the scope of the project. It could include the measured use and time savings if the investment has been completed.

The resulting changes from the control values, as printed from REMI Policy Insight, are shown on Figure 6.14 and Table 6.1 as difference and in Table 6.2 and Table 6.3 as percentage changes (see Figure 6.14 and Tables 6.1, 6.2, 6.3 below).

Figure 6.14. **Sample demo model simulations**
Reduction of internal transport and access costs by 5%
Analytical organisation (use graph option):
Graph 4: Output components by demand source
Differences as compared to REMI standard reg control

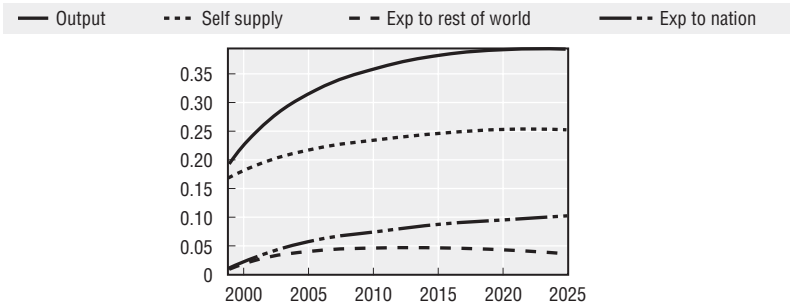


Table 6.1. **REMI standard reg control**
Reduction of internal transport and access costs by 5%
Graph 4: Output components by demand source – differences

Variable	1998	2000	2010	2025
Output (1 000 M 97 £)	0.1917	0.2466	0.3597	0.3924
Self supply (1 000 M 97 £)	0.1691	0.1887	0.2353	0.2525
Exp to multiregions (1 000 M 97 £)	0	0	0	0
Exp to rest of nation (1 000 M 97 £)	0.01009	0.0258	0.04763	0.0378
Exp to rest of world (1 000 M 97 £)	0.01246	0.03203	0.07679	0.1021

In Figure 6.14 and Tables 6.2 and 6.3 three components of output change (export multi-region are unchanged because this is a one region model). The exports to the rest of the world and to the rest of the nation increase because the cost of production decreases relative to its value in the calibration year.

The self-supply (on Figure 6.14 and all Tables) increases because the share of the Demo area supplied locally increases (the regional purchase coefficient, Table 6.3) and the demand (Table 6.3) increases. The local share increases because relative production cost decreases (Table 6.3). The demand increases because consumption (not shown) increases due to the increase in real disposable income increases (Table 6.3) and the increase in the regional share supplied locally (The Regional Purchase Coefficient). The demand also

Table 6.2. **REMI standard reg control**
Reduction of internal transport and access costs by 5%
Graph 4: Output components by demand source – per cent change

Variable	1998	2000	2010	2025
Output (1 000 M 97 £)	0.74%	0.93%	1.26%	1.39%
Self supply (1 000 M 97 £)	2.10%	2.26%	2.49%	2.60%
Exp to multiregions (1 000 M 97 £)	n.a.	n.a.	n.a.	n.a.
Exp to rest of nation (1 000 M 97 £)	0.09%	0.23%	0.45%	0.47%
Exp to rest of world (1 000 M 97 £)	0.18%	0.45%	0.89%	0.98%

Table 6.3. **REMI standard reg control**
Reduction of internal transport and access costs by 5%
[Top] – per cent change

Variable	1998	2000	2010	2025
Employment (thous)	1.23%	1.39%	1.61%	1.69%
GRP (1 000 M 97 £)	0.83%	1.04%	1.40%	1.54%
Pers inc (1 000 M Nom £)	0.47%	0.59%	0.81%	0.96%
PCE-price index (UK 1997 = 100)	–0.54%	–0.54%	–0.55%	–0.56%
Real disp pers Inc (1 000 M 97 £)	0.83%	0.92%	1.11%	1.30%
Population (thous)	0.01%	0.04%	0.21%	0.49%
Econ migrants	1.56%	1.58%	2.10%	2.29%
Total migrants	1.56%	1.58%	2.10%	2.29%
Labor force	0.17%	0.34%	0.71%	1.00%
Demand (1 000 M 97 £)	0.74%	0.90%	1.11%	1.21%
Output (1 000 M 97 £)	0.74%	0.93%	1.26%	1.39%
Delivered price	–0.48%	–0.48%	–0.49%	–0.50%
Rel cost of production	–0.26%	–0.26%	–0.27%	–0.28%
Labor intensity	0.00%	–0.01%	–0.05%	–0.09%
Labor access index	0.01%	0.03%	0.06%	0.07%
Indust mix index	0.00%	0.00%	0.00%	0.00%
Reg pur coeff (SS over Dem)	1.35%	1.35%	1.36%	1.38%
Imports (1 000 M 97 £)	–0.17%	–0.02%	0.13%	0.19%
Self supply (1 000 M 97 £)	2.10%	2.26%	2.49%	2.60%
Exports to multiregions (1 000 M 97 £)	n.a.	n.a.	n.a.	n.a.
Exports to rest of nation (1 000 M 97 £)	0.09%	0.23%	0.45%	0.47%
Exp to rest of world (1 000 M 97 £)	0.18%	0.45%	0.89%	0.98%
Wage rate (Thous nom £)	–0.13%	–0.03%	0.15%	0.15%

increases because of increase in investment (not shown) and local government spending (not shown). Real disposable income increases occur because employment (Table 6.3) is up and prices are down (Table 6.3).

Table 6.3 shows the average nominal wage rate down slightly in the first year even though the wage rate in every industry increases (not shown). This happens because the lower wage industries (services, etc.) increased their employment at a greater rate than the higher paying industries. The increase in the labor access index, which increases productivity, is due to the expanding economy and number of employees. Note also the GDP per capita increase in the short and long term (GRP and population, Table 6.3).

In addition to the information shown, over several thousand other variable values are accessible. The tables and graphs include information on all of the concepts used by industry and by age cohorts. Using the information shown, as well as the information from the other REMI graphs and tables makes it possible to fully document and explain the effects of structural fund investment.

Policy recommendations

Local and central authorities use economic models to make more effective decisions. Governmental authorities have limited resources to apply to economic development purposes. Therefore, economic models and other analytical tools are important because they provide a means of evaluating alternative uses of resources.

We suggest that decision makers use models that have six key features: 1) specification to local conditions, 2) strong theoretical and structural foundation, 3) integrated general equilibrium, input-output, econometric, and economic geography methods, 4) a comprehensive set of both input and output variables, 5) year-by-year results, 6) a record of use for a large range of projects over many different regions.

The calibration of the model to a specific economy should incorporate local data. However, econometric parameters in the model should be based on a large set of cross-section and time-series data. The larger data set should be used for estimation purposes, since statistically robust estimates for a structurally well-defined model may not be possible using time series data for a single region.

Policy analysis should be clear and fully defensible. A model that is based on a strong theoretical and structural basis will incorporate well-defined cause-and-effect relationships in the economy. The basis of policy recommendations from such a model will be fully transparent, enabling policy makers to be fully informed.

An integrated modeling approach brings together the advantage of differing methodologies in a comprehensive and consistent system. Input-output structures are important in tracking inter-industry relationships in the economy. General equilibrium responses capture important long-term

responses to price, cost, and wage signals. Econometric techniques validate the empirical basis of a model, and new economic geography methods explain how agglomeration economies are significant in understanding the dynamics of regional development.

A comprehensive model framework, including detailed policy variables for all parts of the economy, is needed in order to adequately represent the types of policies that may be proposed for economic development or other purposes. Models that are developed with a single purpose or application may be quite limiting. For example, an economic development plan may include a worker training aspect; thus, a model should incorporate policy variables that allow the user to change labor productivity in order to fully capture the effects of the proposal.

For both planning and analysis purposes, a dynamic model with year-by-year results is critical because it provides for a clear understanding of program implications over time. A medium- and long-term economic forecast is invaluable for the determination of needs for highways, airports, power generation plants, schools, and other investments. Many policies will also have significant long-term implications; for example, many of the economic benefits of new transportation investments occur for decades after the initial construction occurs. Analysts also need to understand the timing of effects in order to evaluate the overall feasibility of a project.

The performance of an economic model should ideally be validated through its application for numerous studies for diverse economies. Although a model can be built and tested in its initial development, the performance in many real-world studies is vital as part of a model development process. Model builders are able to continually enhance and refine an economic model as a result of the experiences of model users.

Notes

1. See Fan, Wei, Treyz and Treyz (2000).
2. See Treyz, Frederick and George Treyz. "The REMI Economic Geography Forecasting and Policy Analysis Model." August 1, 2001.
3. See Treyz, George and Lisa Petraglia (2001)
4. See Rickman, Dan S., Gang Shao and George I. Treyz (1992).
5. See Greenwood, *et al.* (1991).
6. See REMI Staff (2002). "REMI Policy Insight, Model Documentation, European Version 5.1", Regional Economic Models, Inc., p. 46-52.
7. Prof. Dr. Michael Wegener, Institute of Spatial Planning (IRPUD), August-Schmidt-Str.6, 44221 Dortmund, Germany.
8. See Weisbrod, *et al.* (2001).

9. See Systematic Solutions, Inc. (1992).
10. See Minnesota Racing Commission (1991).
11. See Lenk, Franklin (1990).
12. See Blois, Tara, *et. al.* (1995); Deloitte and Touche LLP (1995); and Sims, Richard (1994).
13. See Kentucky Legislative Research Committee (1986).
14. See Fulton, Grimes and Baum (1984).
15. See Boston Redevelopment Authority (2000) and Connecticut Center for Economic Analysis (1992).
16. See Connecticut Center for Economic Analysis (1993) and Deller, Steven C., *et al.* (1992).
17. See Warner, *et al.* (1996).
18. See Bonardelli, Mark A. (1995).
19. See Bartlett, *et al.* (2001).
20. See Otto, D. and M. Lipsman (1993).
21. See University of Connecticut (1994).
22. See Nexus Associates, Inc. (1995).
23. See Harkenrider, Greg (1999).
24. See Wilbur Smith Associates (1992).
25. See Economic Research Associates (1996).
26. See Smith, Anne E., *et al.* (1997).
27. See Myers, J.G., C.A. Pasurka, Jr., and T. Veselka (1987).
28. See ICF, Inc. (1993).
29. See Duncombe, William, *et al.* (1997).
30. See Weiskoff, Richard (2000) and Weiskoff, Richard (2002).
31. See Silkman, Richard (1987).
32. See Vermont Department of Public Service (1991).
33. See Hickman, James E. (1995) and Weisbrod, *et al.* (1995).
34. See Wyoming Public Service Commission (1996).
35. See Baum, A.L., G. Fulton and D.R. Grimes (1984).

References

BAUM, A.L., G. FULTON and D.R. GRIMES, "Industrial Location Decisions and Their Impact on the Michigan Economy: The Mazda Automobile Assembly Case", 1984.

BLOIS, Tara, Steven R. CUNNINGHAM and William F. LOTT, "The Bridgeport Casino Proposals." Connecticut Center for Economic Analysis, 42 pages; October, 1995.

BONARDELLI, Mark A., "Analyzing the Impact of Lost Coal Sales Using the Illinois REMI Model". *International Journal of Public Administration*, Vol. 18, No. 1, pp. 101-118; 1995.

BOSTON REDEVELOPMENT AUTHORITY POLICY DEVELOPMENT AND RESEARCH DEPARTMENT, "An Economic Analysis of Fenway Park and the New Fenway Park Proposal (Year 2000)". April 13, 2000.

CHRISTOPHER, JR., Chris G., CHENGFENG LOU and George I. TREYZ, "Regional Labor Force Participation Rates". March 3, 1996.

CONNECTICUT CENTER FOR ECONOMIC ANALYSIS, "New England Patriots Franchise Acquisition: An Economic Impact Study". 41 pages; November 12, 1992.

DELOITTE and TOUCHE LLP, "Economic Impacts of Casino Gaming on the State of Michigan". 114 pages; February, 1995.

DUNCOMBE, William, Shannon FELT, James R. FOLLAIN, and Bernard JUMP, Jr., "The Economic and Fiscal Impact of Lake Remediation on Onondaga County". Center for Policy Research; Metropolitan Studies Program Series Occasional Paper No. 186; June 1997.

ECONOMIC RESEARCH ASSOCIATES, "Economic Impact and Benefit/Cost of High Speed Rail for California (final report)". 1996, 154 pages.

FAN, Wei, Frederick TREYZ and George TREYZ, "An Evolutionary New Economic Geography Model", *Journal of Regional Science*, Vol. 40(4), 671-695, 2000.

FUJITA, Masahisa, Paul KRUGMAN and Anthony J. VENABLES, *The Spatial Economy: Cities, Regions, and International Trade*. Cambridge, Massachusetts: MIT Press, 1999.

FULTON, George, Personal communication, November 1, 2002.

FULTON, G., D.R. GRIMES and A.L. BAUM, "Industrial Location Decisions and Their Impact on the Michigan Economy: The Mazda Automobile Assembly Case". 58 pages; November 1984.

GREENWOOD, Michael J., Gary L. HUNT, Dan S. RICKMAN and George I. TREYZ, "Migration, Regional Equilibrium, and the Estimation of Compensating Differentials". *American Economic Review*, 1991.

HARKENRIDER, Greg, "Prison Study Economic Report", Commonwealth of Kentucky. 17 pages; September, 1999.

HICKMAN, James E., "Economic Opportunities Through Energy Efficiency: An Alternative Analysis". 12 pages; January, 1995.

ICF, INC., "Economic Effects of Adopting a Low Emission Vehicle Program in Maryland". Prepared for the American Lung Association of Maryland, Inc. February, 1993.

KENTUCKY LEGISLATIVE RESEARCH COMMITTEE, "Economic and Fiscal Effects of the Toyota Auto Facility on the Kentucky Economy". 13 pages; 1986.

LENK, Franklin, "The Economic Impact of Expanding Bartle Hall and Building an 800-Room Hotel". 26 pages; February 6, 1990.

OTTO, D. and M. LIPSMAN, "Economic Impacts of the 1993 Iowa Floods". 6 pages; 1993.

MINNESOTA RACING COMMISSION, "The Economic Impact of the Horse Racing and Breeding Industry on the State of Minnesota". 9 pages; April, 1991.

MYERS, J.G., C.A. PASURKA, Jr. and T. VESELKA, "An Input-Output Simulation of the Impact of Acid Rain Legislation on Illinois". August 1987; 37 pages.

NEXUS ASSOCIATES, INC., "Executive Summary: The Impact of Tufts University School of Veterinary Medicine on the Massachusetts Economy". 29 pages; October, 1995.

REMI STAFF, 2002, "REMI Policy Insight, Model Documentation, European Version 5.1", Regional Economic Models, Inc.

REMI STAFF,2002, "REMI Policy Insight, User Guide, Version 5.1", Regional Economic Models, Inc.

RICKMAN, Dan S., Gang SHAO and George I. TREYZ, "Multi-regional Stock Adjustment Equations of Residential and Non-residential Investment in Structures", *Journal of Regional Science*, 1993.

RICKMAN, Dan S., Gang SHAO and George I. TREYZ, "The REMI Economic-Demographic Forecasting and Simulation Model", *International Regional Science Review,* 1992.

SILKMAN, Richard, "The Effects of a Mandatory Early Shutdown of Maine Yankee". 38 pages with 52 pages of appendices; September, 1987.

SIMS, Richard, "Economic Impact Analysis: The Lucky Spa Casino". Office of Economic and Tax Policy, Bureau of Legislative Research. 5 pages; 1994.

SMITH, Anne E., *et. al.,* "Costs, Economic Impacts, and Benefits of EPA's Ozone and Particulate Standards". Reason Public Policy Institute/Decision Focus Inc., Policy Study 226; June, 1997; 114 pages.

SYSTEMATIC SOLUTIONS, INC., "REMI Interface with Energy 2020". 18 pages, 1992.

TREYZ, Frederick and Jim BUMGARDNER, "Monopolistic Competition Estimates of Interregional Trade Flows in Services". *Regional Cohesion and Competition in the Age of Globalization,* 2000**.**

TREYZ, Frederick and George TREYZ, "The REMI Economic Geography Forecasting and Policy Analysis Model". August 1, 2001.

TREYZ, Frederick and George I. TREYZ, "The REMI Multi-regional US Policy Analysis Model". Annual North American Meeting of the Regional Science Association, 1997.

TREYZ, George and Lisa PETRAGLIA, "Consumption Equations for a Multi-regional Forecasting and Policy Analysis Model". *Regional Science Perspectives in Economic Analysis.* Eds. Michael Lahr and Ronald Miller. Elselvier Science B.V., 287-300, 2001.

TREYZ, George I., "Policy Analysis Applications of REMI Economic Forecasting and Simulation Models". *International Journal of Public Administration*, 1995.

UNIVERSITY OF CONNECTICUT, "1995 World Special Olympics – Economic Impact Analysis". 11 pages; April, 1994.

VERMONT DEPARTMENT OF PUBLIC SERVICE, "Vermont Comprehensive Energy Plan". 203 pages; January, 1991.

WARNER, K.E., G.A. FULTON, P. NICOLAS and D. GRIMES, "Employment Implications of Declining Tobacco Product Sales for the Regional Economies of the United States". *The Journal of the American Medical Association*. Vol. 275, No. 16, p. 1241-46; April 24, 1996.

WEISBROD, Glen of Hagler Bailly Consulting, INC. Karen POLENSKE and Teresa LYNCH of MIT and Xianuan LIN of Boston University, "Final Report: The Economic Impact of Energy Efficiency Programs and Renewable Power for Iowa".

WEISBROD, Glen, Donald VARY and GeorgE TREYZ, "Project 2-21 – Final Report: Economic Implications of Congestion". *Transportation Research Board, National*

Cooperative Highway Research Program Report 463. Washington, DC: National Academy Press, 2001.

WEISKOFF, Richard, "Missing Pieces in Ecosystem Restoration: The Case of the Florida Everglades". *Economic Systems Research*, Vol. 12, No. 3, 2000.

WEISKOFF, Richard, "The Economics of Everglades Restoration: A Tale to Two Models". School of International Studies, University of Miami; 16 pages; 2002.

WILBUR SMITH ASSOCIATES, "Executive Summary: US Highway 20 Corridor Development Study". 18 pages; 1992.

WYOMING PUBLIC SERVICE COMMISSION, "Electric Utility Industry Restructuring Issues". Summary of six Wyoming stakeholder subcommittees. 61 pages; November 12, 1996.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 7

A Commentary on Frederick and George Treyz's Paper and the Workshop "Analysis Policies for Local Development Using Forecasting Models"

by
Robert Wilson,
Principal Research Fellow Institute For Employment Research,
Warwick University, United Kingdom

The presentation

The workshop began with a brief introduction to what REMI can do, in terms of the issues it can be used to address and how this is accomplished. Dr Treyz provided a description of the key elements of the model:

- an input output structure;
- general equilibrium elements, with long-term responses to prices and wages;
- an econometric component based on time series modelling to incorporate dynamic effects;
- new economic geography, features including agglomeration and clustering effects.

This was followed by examples of the inputs used to drive a typical model scenario and the outputs that can then be reviewed. These focused on the demands for goods and services, the effects on wage and price changes, and demographic and labour market effects.

Examples of the policy insights which can be derived were then summarised for the US (with examples in a number of different contexts) and also in other countries.

The pros and cons of a model-based approach were discussed, both in Dr. Treyz's presentation and in subsequent discussion. The following key points were made:

- it was argued that macroeconomic insights are essential for rational policy decision making. micro-studies take us only part of the way to a complete evaluation;
- a quantitative macro-model can provide a common framework for comparing the overall fiscal and employment effects of alternative policy interventions;
- such an all encompassing and comprehensive approach can also deal with displacement effects across geographical boundaries;
- such a general model enables the incorporation of theoretical insights supported by empirical testing from a variety of disciplines;
- the model framework facilitates transparency in terms of assumptions and implications;
- the dynamic element of the model framework is essential to evaluating when effects take place.

The need to regard the use of such methods as a process, with feedback from users and subsequent modifications influencing the final outcomes, was emphasised. Such analysis should not just be a "one off" exercise.

If done appropriately, this can help to empower users and ensure involvement and ownership of the results and outcomes.

Alternative macro-econometric models

Other macro econometric models have also been developed which can be used to help evaluate labour market interventions. In the United Kingdom, for example, the Institute for Employment Research (IER), in collaboration with Cambridge Econometrics (CE), have developed the Local Economy Forecasting Model (LEFM). As with REMI, LEFM is based on an assumption of common behavioural patterns and technical linkages applying at national, regional and local levels, while recognising the structural differences that make each geographical area unique.

LEFM is a tailored software tool that provides local economy analysis and forecasting in the United Kingdom. Since its inception in 1993 this has been set up for hundreds of areas in the UK, as well as a few in mainland Europe. Its prime function is to guide policy makers and analysts on fundamental economic and labour market trends at a local level.

LEFM has been designed to fulfil the following criteria:

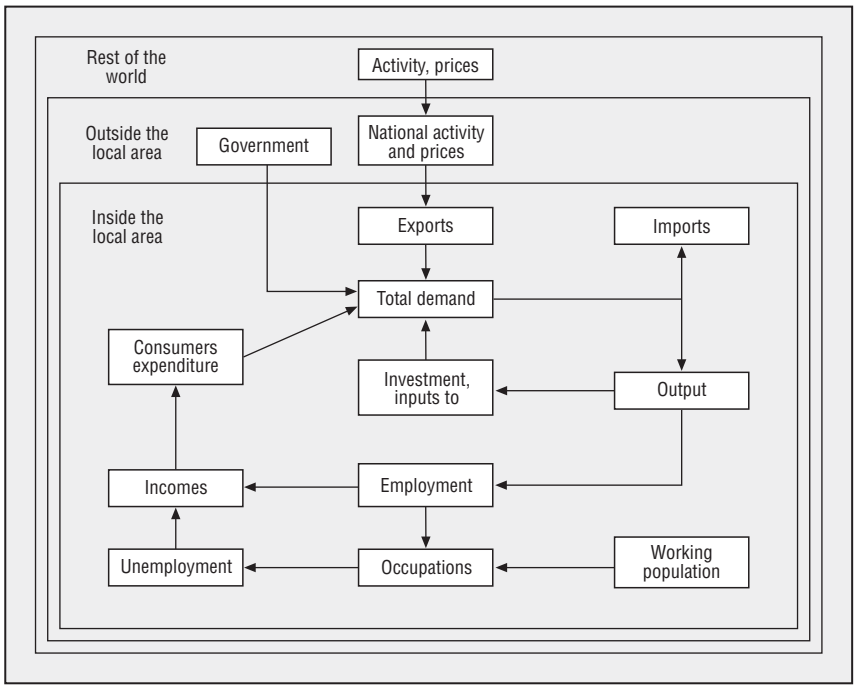
- an efficient means of generating tailored local economy projections that makes maximum use of the national, regional and local information available;
- easy updating, allowing the user to draw on and reassess previous analysis;
- a rigorous and transparent method of analysis, yielding results that can be readily traced back to assumptions;
- easy links to regularly updated, authoritative forecasts at the national and regional levels;
- an explicit way to introduce local knowledge and views;
- substantial sectoral detail, so that projections can be identified closely with major local firms;
- labour market detail (sector, gender, occupation);
- the ability to implement scenarios and sensitivity testing;
- the ability to carry out impact studies (*e.g.* opening/closure of a large establishment);
- easy access to results for evaluation, plotting and file transfer to other software for report writing, presentations, etc.

In LEFM employment data play a central role, since it is on this variable that the most detailed and reliable information is available at local level. Most of the economic indicators at local level are derived by combining information on the corresponding regional level indicator and the local area's share of regional employment or population.

The model's basic structure (simplified) shares a number of features in common with REMI (see Figure 7.1). It is based around a detailed Leontieff input output table. It provides a dynamic solution using annual historical data from 1981 with prospectus forward to 2015. However, in contrast to REMI, LEFM is focussed primarily upon the demand side. LEFM distinguishes 50 employing activities and 6 status/gender types. In addition, the occupational results provide a breakdown by 2 occupational categories, for males and females separately. It is comprised of a series of behavioural and technical relationships at local level which mirror those at national and regional level in terms of parameter values, etc. The main elements are:

- Consumer demand.
- Investment demand.

Figure 7.1. **LEFM – A local economy forecasting model**



- Government expenditure.
- Exports and imports.
- Intermediate demand.
- Total commodity demand (Q).
- Industrial gross output (Y).
- Value o.
- Employment ondustry (YEO).
- Employment by gender and type.
- Employment by occupation.
- Labour supply, etc.
- Unemployment.
- Incomes.

LEFM is calibrated so that employment in the local area follows the observed historical patterns, while the forecast moves in line with regional or national trends. This is based on a simple econometric analysis. Calibration is achieved by altering local gross output levels. This in turn is achieved by amending the export share ratio. This ensures that, over the historical period, the model tracks the observed employment data and that the initial forecast is "sensible". The adjustments to the "net commuting" residual also ensures that the model traces the observed unemployment data and that the initial base forecast changes in line with more general regional trends. Complete details of the model can be found in Wilson *et al*. (1995).

Initial discussion: statistical significance and validation

The statistical significant and robustness of the results from REMI were queried and questions were also raised about the applicability of US parameters to European contexts. This can be seen as part of a more general debate about the applicability of model results based on particular historical experiences to other contexts and periods.

Dr Treyz emphasised that considerable care was taken to ensure that estimated parameters were statistically robust, but recognised that, when looking at forecast results, it is much more difficult to make precise statements about statistical accuracy. Rather, the emphasis should be on producing reasonable or sensible outcomes and focusing on the difference between policy "on" and "off" or differences between different policies.

There was a lengthy discussion on the problems of applying REMI in a European context, given the different social and economic context. As well as problems of data classification and consistency, this assumes that the performance and behaviour are common across international boundaries.

While this may be true in a very general sense, different institutional and legal frameworks undoubtedly affect the way people behave, the incentives they face, etc.

The social context of a given situation, social cohesion and similar factors are not readily incorporated into a quantitative model framework. The general parameters should be regarded as a starting point subject to modification to reflect local circumstances. This reinforces the importance of getting local involvement when applying these tools, so that such local factors can be taken into account. Models such as REMI and LEFM enable local evaluation analysts and policy makers to gain insights based on general findings from economic analysis and from other disciplines.

There was some discussion of the complexities of using such models, which does undoubtedly require a certain minimum level of expertise. The results from applying such approaches can be quite sensitive to how they are operated and the input assumptions used. As always there is the danger of "garbage in garbage out".

A further concern was raised about the timeliness of some key data inputs, especially elements such as input output tables. However, such values do tend to change quite slowly over time, so this was not felt to be a major issue. This again emphasises the importance of focusing on relative effects (of one policy compared to another, policy on compared to policy off) rather than absolute changes.

Usefulness of macro models

Earlier sessions in the conference, particularly the contribution by Bartik, emphasised the importance, when evaluating particular policies and programmes, of moving beyond the immediate impact on individuals or firms. In particular, a comprehensive evaluation requires an assessment of the overall effect on revenues and total employment levels. While the various approaches to evaluation advocated in previous sessions focused on issues such as deadweight and displacement at a micro level, they generally provide no mechanism for reaching a macroeconomic overview of the effects. Equally, while these other methods provide various means by which the past impact of such possible interventions may be assessed they provide no insight into possible future effects.

In order to obtain these additional insights a macroeconomic forecasting model is needed. This provides the mechanism by which the macroeconomic consequences of individual effects can be summed (building in multiplier and input-output linkage effects) and also providing a useful counterfactual of what might have happened in the absence of the policy intervention. The key conclusion from the discussion is that such models, while not a panacea, and

while providing just one leg of a complete evaluation approach, can deliver important insights into the evaluation process. They help to remind us of all the complex interactions that should be taken into account but which are sometimes forgotten.

Other papers in the conference (*e.g.* the contribution by Wong) emphasised the need to take into account geographical displacement effects. Macro models cannot by themselves answer this question but they provide an important framework within which it can be addressed.

Bartik also emphasised that rigorous evaluation is possible through a variety of methods but including the need to link in regional econometrics with fiscal impact and local labour market models. A mixture of techniques, quantitative and qualitative, is needed with a focus on improving programmes rather than "killing" poor ones.

Smith also raised the importance of synergy and macroeconomic effects which might otherwise be neglected in a purely microeconomic evaluation.

Oldsman emphasised the importance of establishing robust baseline data. A key aspect of this is providing a counter-factual – what would have happened anyway, even if the policy had not been introduced/changed?

Eberts and O'Leary made the important point that social programmes that are intended to affect local and regional economies are especially difficult to evaluate. Their scope is much broader than simply looking at impacts on individual actors or organisations. On the other hand, the scale of interventions are often so small as to have indiscernible effects. A macro modelling tool provides a means to at least get a handle on these kinds of outcomes.

Walker emphasised the need to collect longitudinal data to facilitate model building and the development of better, more sophisticated forecasting tools.

Finally, another important issue is that polices and subsequent evaluations may interact and evolve, changing behaviour and generating additional capacity (Stern). This requires quite careful handling in the model framework, in which the parameters are based on previous patterns of behaviour and institutional and related frameworks.

Wong reminded us that intervention in the social sciences is rarely a matter of repeating a controlled experiment. Often the carrying out of the experiment, as well as any evaluation of it, will influence the outcomes and behaviours of those involved. This applies with special force when using macro models of the REMI or LEFM type. Changing behaviour may result in the need to reassess the explicit assumption of fixed parameters based on historical experience. Robson emphasised a similar point, citing the importance of action orientated research, perceptual indicators and softer, qualitative, rather than hard quantitative approaches.

Reference

WILSON, R., A. ASSEFA and J. BEARD (1995), "A Local Economy Forecasting Model (LEFM) for the UK Economy". Paper presented at the European Symposium of Labour Market Developments, University of Warwick, 18th-19th May, 1995.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 8

Area-based Policy Evaluation

by
Brian Robson,
Centre for Urban Policy Studies,
University of Manchester,
United Kingdom

The policy context

Formal urban policies in Britain date from the late 1960s and, over the last three decades, policy has placed great faith in area-based initiatives (ABIs) as a mechanism to tackle the problems associated with dereliction and deprivation in the large ex-industrial towns and cities of Britain (for example, Hall and Nevin, 1999).

Since the 1980s there has been a bewildering array of different types of policy instrument, amongst which have been (or are):

- Action for Cities (AfC) in the 1980s which focused additional resources on 57 local authority districts (LADs).
- City Challenge (CC) which operated between 1992 and 1998 and targeted resources at 31 sub-district areas.
- the Single Regeneration Budget (SRB) which rolled together the expenditures from 21 earlier programmes, went through 6 rounds of allocation during the 1990s, and was largely targeted at small geographically defined areas which received funding for periods of up to seven years.
- New Deal for Communities (NDC) which currently supports 39 sub-district areas with populations of up to 4 000 over a period of ten years.
- Urban Regeneration Companies (URC), the first of which were created in 1999, with 11 having now been established to develop master plans mainly for city centre areas, and with running costs funded through partnerships between Regional Development Agencies (RDAs), LADs and English Partnerships (EP).

Each of these programmes has been funded by the government department with principal responsibility for urban regeneration – successively named the Department of Environment (DoE), Department of Environment Transport and the Regions (DETR), Department of Transport, Local Government and the Regions (DTLR), and now the Office of the Deputy Prime Minister (ODPM).

In addition, in the last five years, there has been an increasing number of complementary area-based initiatives from other central government departments. Examples include the Crime Reduction Programme, Sure Start, Education Action Zones, Employment Zones, Sports Action Zones, and Health Action Zones – from government departments such as the Home Office, Employment, Education, Sport and Culture, and Health.

Such area-based approaches are predicated on the belief either that there are area-related processes that compound the problems that are faced by deprived individuals, or that there are efficiencies associated with the delivery of policy within defined targeted areas. While there has been much agonising in the academic literature about whether or not an "area effect" can be demonstrated (Dorling *et al*., 2001; Atkinson and Kintrea, 2001), there seems to be general agreement that there are administrative benefits associated with spatial targeting – not least the potential synergy that can be achieved across different policy domains, and the efficiency of deploying personnel and resources within a limited number of areas. The case for area-targeting is that, given the wide disparities in deprivation, the neighbourhood is the most appropriate scale for fostering community identity and involvement and that resources concentrated at small areas over a number of years can achieve a step improvement in the circumstances of deprived areas (Lawless *et al*., 2000).

Over time, the focus of ABIs has changed. At the outset, in the 1970s, most programmes broadly covered economic, social and environmental objectives, but the additional targeted resources were generally small. In the 1980s, the majority of funding was directed to the physical improvement of derelict areas in an attempt to revive the working of property markets. Typical interventions included programmes such as Enterprise Zones which offered financial incentives to firms to locate in decayed urban sub-areas, or the Urban Development Corporations which were run by private-sector boards with the aim of re-furbishing the infrastructure of derelict ex-industrial sites. In the 1990s, City Challenge and the SRB programmes brought local authorities back into the frame and broadened the focus to encompass social and economic issues as well as environmental and property-related aims. With this change of emphasis, two principal features have come to dominate recent approaches to urban regeneration; a stress on partnership working and, related to this, the aim of developing better co-ordination across policy domains.

The recent Urban White Paper (DETR, 2000) fundamentally changed the main thrust of policy. It moved away from the previous almost exclusive dependence on area-based initiatives and espoused the aim of "mainstreaming" as a way of better tackling urban problems. Mainstreaming can be thought of in three distinct ways:

- the attempt to bend resources from main spending programmes (such as education, social services, housing) to target areas of especial need or to improve the quality of service delivery to such areas;
- the attempt to learn lessons from what works in specific programmes and projects and apply them more generally to other areas; and
- the attempt to incorporate into mainstream services the policy lessons that arise from specific initiatives.

While mainstreaming has long been a mantra within urban policy, the new policy framework is the first to put it centre stage. Local Strategic Partnerships are now in process of being established in local authorities and they are charged with developing Community Strategies that consciously use mainstream as well as specific resources from funds such as the Neighbourhood Renewal Fund (which is targeted at 88 local authorities) to reverse the fortunes of deprived areas.

So, even though many area-based initiatives remain – not least the NDCs, URCs and some of the later rounds of SRB funding – English urban policy has now drawn a line under its almost exclusive dependence on area-based initiatives.

The growth of monitoring and evaluation

From the outset of this 30-year history of urban policy, there has been a continuous development of monitoring and evaluation under the auspices of central government. The total resources channelled at monitoring and evaluation have grown considerably. On the face of it, evaluation has now become an integral part of the policy environment. As well as the evaluation projects sponsored by funding departments of central government, generic policy reviews are undertaken by bodies such as the Audit Commission (for example, Audit Commission, 1989; 2002) and the National Audit Office; and the newly-established co-ordinating units established by central government such as the Social Exclusion Unit and the Regional Co-ordination Unit also draw on research-based evidence with the aim of steering the direction of policy. A recent example is the Regional Co-ordination Unit's review of area-based initiatives (RCU, 2002) which makes recommendations about merging and mainstreaming several of the existing separately funded regeneration programmes from across a range of government departments. In addition to such government-sponsored evaluation, there have been programmes evaluating area-based policies both by research charities – especially the Area Regeneration Programme of the Joseph Rowntree Foundation (see, for example, Maclennan, 2000) – and by the Economic and Social Research Council through its Cities Programme (see, for example, Begg, 2002). In this paper, however, the focus is restricted more narrowly to government-sponsored research.

Much of the research supported by the Department of the Environment in the 1970s and 1980s was relatively haphazard. Most of the evaluations were *ex post* and researchers had some difficulty in assembling data that could identify the initial conditions when programmes started and in recreating data on outputs during the lifespan of initiatives. However, since the 1990s (and especially since City Challenge and the SRB programmes), most funding of policy interventions has usually been contingent on continuous monitoring

and evaluation by local partnerships themselves and this has been accompanied by national evaluations undertaken for ODPM. The step change in evaluation came with the advent of the new government in 1997. There is now a firmly-embedded culture of local and national evaluations in virtually every area-based initiative.

This growth in evaluation is a reflection of the government's growing emphasis on seeking "value for money" and on its mantra of "what matters is what works". The notion of evidence-based policy-making has been most clearly seen in the field of medicine (where, for example, the Cochrane Collaboration and units such as the Centre for Evidence-Based Medicine in Oxford have developed a range of reviews of evidence-based studies of health care). Translating such experience into assessing the impacts of area-based regeneration has offered some major challenges to the social sciences.

Evaluation in the 1970s to 1990s

To the cynical eye, the official urge to monitor and evaluate quickly became institutionalised and routinised and thereby lost its cutting edge. For example, Ho has suggested a three-fold categorisation over the period between the 1970s and 1990s – what she calls three ages of official evaluation research: "innocence" in the early years of the 1970s; "dissent" in the 1980s; and "acquiescence" in the 1990s (Ho, 1999). Rather than closing the loop of monitoring/evaluation/policy reformulation, she argues that, over time, evaluation was increasingly used merely to justify and applaud what had been done.

Three successive evaluation projects can be used to illustrate this shifting balance of types of approach.

Community Development Project

The first is most strikingly illustrated by the formative evaluations of the Community Development Project (CDP), a programme which saw a strong emphasis on action research whose avowed aim was to influence the formulation of policy. The twelve CDP projects established by the Home Office in the 1970s were a major experiment to improve social services for those most in need. Action teams were employed by the respective local authorities, with funding from the Home Office, along with a central Information and Intelligence Unit (Higgins *et al.* 1983). Each team produced in-depth studies of their project areas and a series of inter-project reports was published which offered diagnoses of the causes of poverty and deprivation. These were highly critical of the small scale and narrow focus of the then government policy, not least that it was based on a social pathology philosophy that assumed that problems were internal to small communities rather than being embedded in

the workings of the broader political and economic context (CDP, 1977). Given the direction taken by the research teams and the critical nature of their reports, it was perhaps no surprise that the Information and Intelligence Unit was closed prematurely in 1976 and that the CDP programme was wound up.

Action for Cities

A second example is DoE's evaluation of the collection of initiatives rolled together under the heading of Action for Cities (AfC) at the end of the 1980s (Robson *et al.*, 1994). This project attempted to develop a methodology to address some of the central conundrums of quantitative evaluation which were summarised as six Cs:

Counterfactual, or deadweight, issues. What would have happened anyway, even in the absence of government intervention.

Contiguity, or displacement, effects. The impacts of policy in targeted areas may have negative or positive effects in adjacent areas.

Confound issues. Outcomes can be the result of many different, often overlapping, initiatives so that it is difficult to attribute change to any one of a multitude of programmes.

Contextual effects. Places start from a position of very different assets and potentials and broader national changes can hence impact on localities in very different ways.

Combinatorial issues. The packages of interventions include a variety of different combinations of programmes – addressed to job creation, physical improvement, crime reduction, health improvement and the like – each of which can have spillover effects on the others. Some combinations prove more effective than others.

Choice effects. The sets of places targeted for specific programmes alter over time and across different programmes, so that it becomes difficult to assign places unambiguously to a policy-off or policy-on category.

The need to take account of these issues was subsequently incorporated into the formal Treasury guidelines for evaluation research (HM Treasury, 1995). The Treasury Guidance has "encouraged agencies and partnerships to be more concise in terms of defining objectives and inputs, more sensitive to the importance of establishing net, rather than gross, outputs, and more equipped to assess value for money" (Lawless *et al.*, 2000). Yet, while the Guidance appears to assume that these issues can readily be tackled in evaluative research, in practice they have continued to represent real conundrums for evaluation methodologies and much of the evaluation still focuses on outputs rather than outcomes and has great difficulties in looking at the inter-relationship between different strands of policy.

Since the AfC research was probably the most ambitious overall evaluation during the 1980s and 1990s, it is worth looking at its approach in some detail. It used both quantitative and qualitative methods. Quantitatively, it looked at the relationship between financial inputs and socio-economic outcomes using regression analysis and measures of spatial concentration. Inputs comprised both the targeted funds of the urban programme and "mainstream" resources channelled to local authorities. Outcomes were measured through five high-level indicators: unemployment, job creation, small-firm creation, house price change, and net migration of the 25-34 year-old cohort. By looking at the relationship of inputs to these high-level outcome indicators, the research made two tacit assumptions: that it was impossible to take a set of "policy-off" and "policy-on" comparisons – not only because policy constantly changes, but because few deprived places had not received one or other form of policy intervention; and that in the complex policy arena of regeneration it was difficult to isolate the impacts of specific policy interventions. Hence, its approach to using a quasi-experimental design was to assume that more resources implied more policy intervention and the null hypothesis was that there would be no relationship between inputs and outcomes.

The analysis was conducted at a variety of spatial scales. At a local authority scale, it looked at outcomes for 123 LADs: the 57 Urban Priority Areas (UPAs) which were recipients of direct targeted funding; 40 "marginal" areas with conditions not dissimilar to the UPAs; and 26 "comparator" districts which did not receive any additional resources. The relationship between the input of resources and changes in socio-economic outcome indicators was used to provide a global measure of the overall impact of policy, to test whether having more resources was linked with absolute and/or relative improvements in circumstances.

At a ward scale it looked at changes in the disparity between poor and non-deprived sub-areas in three conurbations; and attempted – rather unsuccessfully – to use multi-level modelling to disentangle relationships at a variety of spatial scales.

In addition, it conducted a range of interviews and questionnaire surveys in three selected conurbations to look at the processes underlying attempts at regeneration. These entailed a large-scale questionnaire of residents, and interviews with employers and policy "experts".

The conclusions suggested that policy had had very mixed results. On one hand, for most of the outcome indicators, there was little relative improvement in the areas targeted by policy. Conditions in the worst areas (especially the cores of conurbations) deteriorated, and the increasing spatial concentration of poverty and unemployment suggested that the level of social exclusion increased. On the other hand, the targeted districts showed relative

improvement in unemployment and in their success in attracting net in-migration of young workers, residents were more optimistic about the prospects of their areas where ABI interventions operated, and there was a positive relationship between the amounts of targeted resource and the relative improvement of areas. Overall, area-based interventions appeared to be beneficial, but not to make significant improvements to the worst areas.

This evaluation falls into Ho's categorisation of "dissent". It recommended a variety of changes to government policy – better co-ordination across programmes, clearer principles behind the targeting of resources, the establishment of an urban "pot" of resources that could be used more flexibly according to local circumstances, more incentives for partnership working across agencies. Many of these were subsequently incorporated into the principles of the Single Regeneration Budget, and to this extent the evaluation can be argued to have been formative as well as summative, whatever government's initial intention had been.

City challenge

The third example is the evaluation of City Challenge which appeared as both an interim (Russell, 1996) and a final report (KPMG, 1999). Both reports focused essentially on outputs rather than outcomes and on the processes involved in the initiative. They analysed output data from the 31 City Challenge areas and conducted a wide range of interviews. As summative evaluations, their accounts were largely based on outputs and structures rather than on outcomes, and to this extent they have something of the flavour of lauding the achievements of a programme which was widely heralded as a valuable break from the narrower focus of the 1980s on physical regeneration and the marginalising of local authorities.

The interim report, for example, developed five main strands:

- compilation of baseline data;
- expenditure and outputs from all 31 of the City Challenge areas, looking at the breakdown of expenditure, sources of funding and outputs in relation to annual targets;
- more detailed case studies in 14 of the areas, drawing on interviews with officers and stakeholders from the public, private and community sectors, together with documentary data;
- a postal questionnaire of key partners in all 31 areas, which included questions on displacement as well as on the structures and processes of the programme; and
- case studies of two areas that had bid for but failed to win City Challenge status.

The final strand was the most novel element of the project. By looking at what had or had not happened in areas which failed to get City Challenge resources, the evaluation was able to draw inferences about the additionality associated with the programme. For example, in one of the unsuccessful authorities, failure to win City Challenge resources meant that other commitments supporting the bid were lost, other deprived areas outside the Challenge area had to be moved further down the local authority list of priorities, and there was a fundamental gap between realising a set of projects (some of which were funded through other targeted resources) and the synergy that would have been achieved with an agreed long-term programme like City Challenge. Each of these points provides some evidence of the additionality associated with the City Challenge programme.

This, and the more direct evaluation of the funded areas, led the team to argue that City Challenge offered a variety of benefits: the incentive to develop more strategic planning; the value of developing flagship projects; the achievement of a critical mass of activity from which linked benefits flowed; and the synergy that could be created across different policy domains.

While both the interim and final reports were largely couched in terms of listing the achievements of City Challenge – the levels of financial leverage achieved, numbers of houses built or improved, jobs preserved or created, derelict land reclaimed or improved, office and industrial floor space created or improved, business start-ups promoted – they also had much to say about the processes and the structures through which the programme was delivered. It would be too harsh a judgement to argue that such evaluations have been merely "acquiescent". As policy has tackled broader inter-connected issues, it is perhaps inevitable that a stronger emphasis has been given to qualitative evaluation. We have learned much about the role of structures, the working of partnerships and the key significance of individuals as a consequence.

There is little doubt that, from the now long history of monitoring and evaluation, successive government administrations have learned much about the challenges of regeneration and about ways in which to develop better approaches to tackle the interlocking elements of urban decay and deprivation. Successive policy changes have reflected this journey up the learning curve, not least in the incorporation of a more community-focused approach to regeneration, the increased emphasis on co-ordination across the different policy domains, and the emphasis on partnership working on the ground.

Recent and current evaluations

Since 1997 there has been a veritable explosion in evaluation. This is reflected in the scale of resources now channelled into evaluation research, with over £8 million devoted to the evaluation of regeneration programmes by

ODPM in the current financial year (Table 8.1). In no small part this can be attributed to the pragmatic non-ideological nature of the new government administration. Its "what works" approach and its increasing stress on the delivery dimension of policy have each helped to encourage a mix of formative and summative evaluations. National evaluations of overall programmes, together with the requirement that local regeneration partnerships conduct their own local monitoring and evaluation, have created a plethora of studies. This has been allied to an approach that has introduced many policy initiatives through initial pilot programmes in selected areas which might subsequently be rolled out more generally to other localities. In addition, the government has provided much wider access to data (through the advent of the National Database of Neighbourhood Statistics) and to the fruits of evaluation projects, most of which have been mounted on government websites.

Table 8.1. **Expenditure by the Office of the Deputy Prime Minister on regeneration programmes and on evaluation research, 2002**

	£
Substantive regeneration programmes	2 366 000 000
Regeneration research	10 000 000
[of which] evaluation research	8 700 000
[of which] evaluation of New Deal for Communities	5 550 000

Source: Private communication, ODPM.

The work that has been spawned from this embodies a wide range of types of evaluation: descriptive, analytical, theoretical, prescriptive and diagnostic. What has been especially evident is the change in the way in which the relationship between research and practice has evolved. Rather than seeing a single continuum from pure research to the development of strategy to changes in practice, there is now a greater realisation of the plurality of the relationship between evaluation and practice. Hence there is now a greater mix of evaluative and action research with teams working more closely with policy-deliverers on the ground. Some of the major evaluations currently underway include:

- The evaluation of the SRB programme (for example, Rhodes *et al.*, 2002). This is a long-term project which has developed quantitative analyses of outputs and outcomes, looked at specific projects that are part of the programmes of local partnerships, and investigated the effectiveness of the partnerships themselves.
- The co-ordination of initiatives in areas with multiple policy interventions (Stewart *et al.*, 2002). This looks at six areas which have been the recipients of almost all of the ABIs and essentially takes the form of action research,

working with the relevant authorities, as well as offering summative views of the outcomes of programmes.

- The 24-cities project (Falk, 2002). This is a qualitative study of 24 selected districts to develop critiques of the "visions" of the local partnerships and to identify examples of good practice that might be transferable elsewhere. It has used visits, workshops with citizens, young people and property groups to identify exemplary practice. It reported its initial findings to the Urban Summit which was held by government in October 2002.

New Deal for Communities

The most ambitious of the current initiatives is the national evaluation of New Deal for Communities which is being undertaken under the auspices of the Neighbourhood Renewal Unit in ODPM. Funding for this work was incorporated from the outset in the national budget of NDC and a series of research teams is being co-ordinated through Sheffield Hallam University. This major project combines elements of formative, summative and fine-tuning evaluation (using the categories suggested by Rossi and Freeman, 1999). Three types of evaluation team are involved:

- partnership teams which are looking at the work of each of the individual NDC partnerships;
- cross-cutting theme teams which are looking at substantive policy domains across all 39 NDC areas; and
- complementary teams looking at technical aspects of the programme.

The partnership teams are each conducting an analysis of the objectives and achievements of the NDC partnerships. This entails a critique of delivery plans, of selected projects, of the structures established in each area, of the degree to which mainstreaming is taking place, and of community involvement in the process of regeneration. Since one of the features of NDC is its attempt to bring communities more firmly on board in determining priorities and in the delivery of regeneration, the success in encouraging resident group participation and in listening to the voices of communities are central features of the evaluations. Hence, the teams have used a variety of interviews with policy makers and policy deliverers and with relevant stakeholders, together with focus groups drawn from resident and community groups. Each of the teams produces evaluations to a common format with templates determined from the centre, as well as more individual interpretations of progress on the ground in the individual NDC areas. The style of work is partly summative and partly formative; the latter taking the form of offering advice and feedback to the partnerships and to ODPM as the sponsoring government department.

The theme teams cover substantive cross-cutting topics such as worklessness, health, crime, and education. They are developing evaluations

through the analysis of primary data from across the 39 NDC areas and of secondary data on outcomes, as well as selecting specific projects from samples of NDCs to identify examples of good practice that might be transferable elsewhere.

The complementary teams comprise three groups: one developing a "traditional" assessment of value for money in terms of expenditure and outcomes; a second team drawing together secondary data (especially on welfare benefit payments) for residents within each of the NDCs; and a third team which will use the forthcoming results of the 2001 census to develop formal baselines of indicators at the start of the NDC programme, with a view to the eventual comparisons that might be made at the end of the NDC programme. In addition, a commercial survey organisation has conducted a major household questionnaire survey within each of the 39 areas to produce data on the attitudes and circumstances of residents. One or more follow-up surveys will be undertaken at later stages in the life-span of the overall programme.

At this stage, the national evaluation is scheduled to continue through to 2005 and there is an expectation that it may cover the entire 10-year life of the NDC programme as a whole.

This national evaluation is being conducted alongside local evaluations which each of the NDC partnerships is obliged to carry out as part of their work. These local evaluations are not being done to a common format, but they entail a mix of household surveys and analysis of outputs and outcomes. In comparison to the national evaluation, they place a greater emphasis on the evaluation of individual projects and less of a focus on outcomes.

Over time, there has clearly been a dramatic growth in evaluation activity in Britain and the nature of the evaluations has changed. While much has been learned about the relevant methodologies, there has been a move away from the more formal quantitative top-down assessments and towards more qualitative bottom-up approaches that have relied on interviews with the deliverers and the recipients of policy, on focus groups and panels, and on social surveys.

Geographical data and targeting

It is clear that one of the major practical challenges faced by all these evaluations has been the need to access appropriate data at relevant spatial scales. Since many policy interventions are targeted at relatively small areas, effective evaluation of outcomes needs small-area data that can be compared over time. Yet, until recently (and with the exception of the decennial census), there has been an absence of such data on which research could draw. Most administrative data held by central government departments have applied to

local authorities – too coarse a scale at which to interpret the impacts of most area-based initiatives. Many of the relevant types of information (for example on housing, aspects of the labour force, or perceived crime) can only be drawn from national surveys whose sample size has been too small to produce robust data even at a district let alone a sub-district scale. Much of the output information has come from local administrative data collected by regeneration agencies themselves and this has the disadvantage that there is no necessary correspondence between the categories used by those who assemble data or across the data-collection methods used by different agencies.

This situation is now changing. Government has established a National Neighbourhood Statistics Database (*neighbourhood.statistics.gov.uk*) which uses local authority wards as its basic framework for reporting data, and is beginning to assemble publicly-available information across a range of relevant data sets. An increasing array of administrative data is now also being assembled on the basis of postcodes which provide a relatively flexible geometry through which to aggregate information to a variety of small-area frameworks. For example, for data on pupils' educational performance, prior to the development of postcode data the results of school examinations had previously been available only on a school-by-school basis. This covered the sequence of "Key Stage" tests taken at primary and secondary school levels as well as formal examinations taken at age 16 (GCSE exams) and at 18 (A-level exams). To look at the performance of pupils in specific areas, researchers were therefore faced with two alternatives. They could either assume that pupils attended the school nearest to them and hence could use school-based data to attribute area-based achievement on a nearest-neighbour principle: a somewhat heroic assumption, even though it is truer for primary than for secondary pupils. Alternatively, researchers could model school-based data by drawing on actual pupil catchment information for samples of schools in order to estimate area-based pupil performance. An example of the latter is the Index of Multiple Deprivation (Noble, 2000). Now, however, the Department for Education and Skills is beginning to assemble data on a pupil-address basis and, once a sequence of data exists, this will enable a far more realistic basis on which to evaluate changes in the educational achievements of pupils.

A second difficulty has been that some government departments use different administrative geographies through which to report data. Health statistics, for example, generally refer to health areas which do not map onto local authority districts. Similarly, reported crime data are generally only available for police beats and police districts which again match only imperfectly onto local authority wards and districts. Widespread use of GIS has been made to resolve such conflicting geographies, but the estimates inevitably lack the robustness of data collected specifically for common spatial areas.

The move to improve the availability of small-area data has come in part from the government's emphasis on a neighbourhood scale in many of its policy initiatives. The National Strategy for Neighbourhood Renewal (Social Exclusion Unit, 2001) has set in train a range of neighbourhood-based interventions to improve the prospects and the management of neighbourhoods. This has helped to focus minds on the need for data at a sub-district scale. It has also been stimulated by the emphasis on the spatial targeting of regeneration resources, where the government has made use of a sequence of indices of deprivation. The first two deprivation indices (Robson, 1995; 1998) were essentially developed at a district scale, but incorporated data for wards (with populations generally less than 10 000) and enumeration districts (with populations generally in the hundreds) from the 1991 census. They introduced a number of innovations: the definition of multiple deprivation as a compound of a small number of "domains", each of which was measured by a range of indicators; the use of chi-square values as a means of standardising scores across different indicators; the production of deprivation scores for nested spatial scales (from ED to ward to district); and a range of measures based on the degree (the summation of values) the extent (the proportion of an area with scores above a cut-off point) and the intensity (the average value of the worst three wards in a district).

The Index of Multiple Deprivation 2000 – the current government index – was developed by a team at Oxford. It built on these innovations, but, importantly, was able to incorporate up-to-date values at a ward scale, not least because of the availability of ward-level data on welfare benefits (Noble, 2000). It uses six "domains" of deprivation – income, employment, health, education, housing and access to services. The methodology used to calculate deprivation scores for wards provides a reflection of data availability at the scale of wards:

The **income** and **employment** measures could be calculated by the Oxford team using direct data on the number of claimants across the range of welfare benefits so as to produce the proportion of the total population who qualified for one or other benefit by virtue of need.

Health includes ward-based data on the recipients of disability benefits and a district-level measure of standardised mortality rates.

Education included a complex modelling procedure to estimate pupil performance on the basis of school-based data.

Housing could only be measured at a ward level by an estimate of unfit houses based on the relatively small national sample of the English House Condition Survey, together with data from the 1991 census.

Access to services (a domain whose values proved negatively related to all other domains) used straight-line distance to a range of services (primary

schools, doctors, post offices and food shops) calculated for recipients of welfare benefits.

The sets of indicators for each of the latter four domains were combined on the basis of weights derived from factor analyses; the domain measures were standardised by ranking the values and using an exponential transformation; and the overall measure was calculated by adding the values of the six domains using predetermined weightings of 25, 25, 15, 15, 10 and 10.

While the Index is the most detailed yet produced, it is clear that, with the exception of benefits data, limitations of data (as well as the methodological difficulties associated with producing composite scores) still make all of such indices problematic. For example, the lack of crime data and of measures of physical dereliction are gaps that are acknowledged by ODPM.

These indices have been extensively used by government to help in targeting resources to areas deemed to be in need. For example, the earlier indices helped in the selection of SRB projects and the IMD guided the allocation of Neighbourhood Renewal Funds.

The indices have also been used to evaluate the effectiveness of spatial targeting (although since they have been one of the determinants of targeting there is a degree of circularity involved). The evidence from the expenditure patterns of the first three rounds of the Single Regeneration Budget suggested that there is a strong positive relationship between expenditure and socio-economic deprivation (Tyler *et al.*, 1998). Of the then 366 local authorities, 30% of funding went to the 20 most deprived, 63% to the 56 most deprived and 81% to the most 99 deprived. The remaining 267 authorities received only 19% of the total funding – normally for small pockets of deprivation within otherwise relatively affluent districts.

A further example of the attempt to develop better small-area information is the work on compiling data on the spatial incidence of expenditure within local authorities (Bramley *et al.*, 1998). This study used three case studies areas – Brent, Liverpool and Nottingham – to analyse locally-relevant public expenditure (social security, health, education, housing, transport, public protection and other local government services) at a ward scale, thereby covering some 70% of total public expenditure. It drew on postcode data from administrative records, survey information on travel to facilities, household surveys to estimate usage rates, GIS apportionment of expenditure, and individual geographical locations for some big capital schemes. They were able to show that there is a wide variation in spending between individual wards, that spending in the most deprived wards is some 45% above that in the least deprived, but that there are significant differences between government departments in their pattern of spending in relation to deprivation. The ability to assemble such expenditure data is clearly critical to

many evaluations of area-based initiatives. The fact that it is possible – although enormously time-consuming and expensive – is at least reassuring for the future evaluation of the impact of area-based regeneration policies and for the assessment of the success of "mainstreaming".

Conceptual problems

A number of problems have continued to pose especial difficulties in this elaborate array of evaluations. One is the difficulty in developing genuinely longitudinal analyses. Almost all of the area-based evaluations are essentially cross-sectional comparisons of areas at two or more points in time. This, of course, ignores the fact that part of any socio-economic change in areas is often the result of household mobility between the two dates, rather than changes in the circumstances of the initial residents. This is a particular problem since regeneration frequently encourages, or indeed is aimed at, attracting new residents. There are innumerable instances where the creation of new jobs taken by local residents prompts those residents to move elsewhere so that, for example, levels of unemployment in the area may stay the same even though some of the previously unemployed residents are now in employment. Equally, physical improvements to an area (not least the building of new houses) often attract new residents who are frequently more affluent, better educated and more likely to be employed than are indigenous households; thereby raising the level of area-based socio-economic indicators even though this may not reflect any improvement in the circumstances of the original residents. Ideally, if the focus of interest is on residents rather than areas *per se*, evaluations should track the changes to initial residents through some form of longitudinal surveys of individuals. For example, the continuing evaluation of the Single Regeneration Budget programme has used repeat social surveys as a valuable approach to measuring change (Rhodes, *et al.*, 2002). This clearly presents considerable logistical problems, especially the challenge of tracking those who have moved out of targeted areas. Attempts have been made to tackle this by using friends and neighbours as sources of information on the whereabouts of those who have moved and through the use of continuous panels of residents (for example in the SRB evaluations of Rhodes *et al.* 2002), but the success rates in tracking out-movers have understandably proved limited. The same difficulties have been experienced with successive panel groups which have attempted to keep the same people involved over successive rounds.

Ironically, Britain is quite rich in longitudinal surveys – examples include the National Child Development Study and the Longitudinal Data which have been collected for samples of identical individuals in successive population censuses since 1981. McCulloch (2001), for example, has used the British Household Panel Study to explore the role of individual *versus* area-based

characteristics in the approach to tackling deprivation. However, for area-based evaluations such national samples are too small to be used to analyse changes at the scale of the specific wards or neighbourhoods to which regeneration programmes apply.

A second difficulty is the accurate estimation of deadweight and displacement. Most studies of deadweight have used self-assessed estimates as a measure by asking businesses or other agencies how far outputs would have happened in the absence of policy intervention (for example, Lenihan, 2001). There are clearly issues of self-interest involved in responses to such questions and the interpretation of the results is therefore problematic. Displacement has generally been assessed through comparisons of control and experimental areas (most notably in terms of the incidence of crime). One innovative approach comes from the evaluation of Urban Development Companies where, using the parallel of housing chains, the broader impacts of new business formation was tracked through identifying what happened to the sites and premises of businesses moving into UDC areas and classifying these as "deaths" or "births" of new firms outside the area of the policy initiative (Robson *et al.*, 1999). Perhaps the classic example of the analysis of displacement is the work done on Enterprise Zones (for example, Tym, 1984) – areas in which fiscal incentives were offered to companies to locate. This suggested that a significant number of inward investments resulted from short-distance moves across boundaries, thereby representing no net gain to the wider city economy.

A third problem is the need to assess outcomes rather than outputs. The difficulty with measuring outcomes is in part a result of the absence of small-area data that can be tracked over time. Ideally, a study of outcomes would need to be able to look at high-level indicators (of measures such as unemployment, poverty, net migration or house prices) at a range of scales from neighbourhoods to city regions and to make comparisons of identical indicators over relatively long periods of time. This would provide the basis for looking not only at substantive outcome changes but also at issues of displacement. As noted above, the absence of such data has until recently made such evaluation extremely difficult. There is also, however, a second difficulty; that the impacts of policy interventions are spread over long periods, often well after the formal conclusion of a specific programme. This is perhaps most true of the health dimension, where significant change might only be expected in subsequent generations (Curtis *et al.*, 2002).

The fourth problem is that of disentangling causality, not least where a variety of policy interventions take place in a limited area. This is clearly a vital component of the question of what works and of assessing value for money. It is probably fair to say that this remains the key conundrum in evaluation. In part it is an operational difficulty; of unpacking the effects of

expenditures from multiple programmes. In part, it is a conceptual problem; of disentangling the lines of causality. Programmes directed at one policy domain may have unexpected consequences in other domains: employment generation may help to reduce crime rates; house improvements may also improve physical and mental health. As regeneration initiatives have become more all-purpose and have simultaneously addressed a whole range of socio-economic problems, so these evaluation conundrums have become greater. However, if the aim of policy is increasingly – and rightly – that of tackling interconnected problems, this strengthens the argument for using high-level outcome indicators to assess broad changes in outcomes and to complement such "black box" assessments with softer evaluations of specific projects to trace the relationships between activities on the ground and effects on social and economic circumstances.

Future directions

Indeed, this blend of process-related evaluations with harder quantitative outcome evaluation seems to be the goal to which the evaluation of ABIs ought now to be aspiring. At present, the more formal "traditional" quantitative approaches to the assessment of outcomes based on quasi-experimental methodologies tend to be in short supply, in comparison to the more qualitative evaluations of processes and structures. In part, this is an inevitable consequence of the emphasis on partnerships as a means of delivering more co-ordinated and sustainable regeneration. This has increasingly prompted evaluations to look at processes – and therefore to use bottom-up qualitative methods – rather than at measurable outcomes. Hence, for example, the approach to looking at the Single Regeneration Budget (Rhodes, 2002) and at Urban Regeneration Companies (Parkinson and Robson, 2000; AMION, 2001) both focused largely on the strength of partnerships. In part, the stress on qualitative evaluation has also been associated with the increasing emphasis on the social dimension of regeneration. Yet, as Armstrong *et al.* (2002) argue, there are still compelling arguments for applying "traditional" quantitative approaches to formal outcome evaluation. Such evaluations have been developed more convincingly in looking at narrowly economic impacts at a regional scale (for example, Moore and Rhodes, 1973) than in evaluating more broadly-based neighbourhood-based initiatives. Armstrong *et al.* argue that such approaches are as relevant to initiatives that have social objectives as to those with purely economic aims. They illustrate this with examples of community economic development (CED) initiatives which have wider social aims in addition to their economic objectives. They show that CED schemes present exactly the same range of conundrums for traditional evaluation methods: multiple objectives; multiple beneficiary groups; the measurement of community capacity building; and effects that

derive from overlapping initiatives. Traditional quantitative approaches still have merits in providing a quantitative top-down view of the effects of programmes with a broader social emphasis.

That said, one of the strongest messages to emerge from the evaluation of regeneration is that it is the less tangible elements that are often key to the achievement of successful sustainable regeneration. Leadership, the quality of key individuals in relevant agencies, sensitivity in handling community-based issues, the learning process in the development of cross-agency partnerships; all of these are vital ingredients to the achievement of long-term sustainable regeneration. And they can best be assessed through sensitive use of interviews, discussions, focus groups, panels and the like. If these are important inputs to the process, it is equally true that a vital element of outcomes is the "feel good" factor, the degree to which local people feel safer, more confident of the prospects of their area, more committed to its future and to the success of regeneration programmes, less cynical. Too much can be made of the hard measurable outputs and outcomes of regeneration: new jobs that may only marginally benefit local people, new buildings and facilities that may not readily be accessible because of entry costs or social frictions, new environments that may rapidly deteriorate if there is insufficient maintenance either by public authorities or private care.

Indeed, two of the probable directions that future evaluation in Britain might take are precisely to address more centrally some of these softer questions. This is evident both in the reappearance of projects which are based on action research, and in the greater attention now being paid to perceptual indicators. A prime example of action research is the two-year project looking at the co-ordination of different initiatives in the six localities of East London, Plymouth, Newcastle, South Yorkshire, Sandwell and West Cumbria (Stewart *et al.*, 2002). It is aimed at supporting the various partnerships responsible for nine different initiatives, many of which focus on overlapping areas. Much of its concern is with encouraging joint working to develop clearer strategies, to encourage the sharing of ideas and information and to re-align service provision in the respective areas. The major evaluation of New Deal for Communities equally incorporates a strong element of action research.

The second thrust is exemplified by the use of large-scale questionnaire surveys in the current NDC evaluation (despite the expense entailed in such work) and in the increasing use of a range of Best Value Performance indicators which government is now assembling as part of its Best Value programme. The latter ask questions of residents about their use of services and about their satisfaction with the quality of services, as well as about the nature of service delivery – for example, the opening hours of libraries, or the accessibility of buildings to the disabled. But, with the growing interest in the

role played by local "social capital", one of the additional challenges that evaluators will need to address is how best to measure this concept. The Home Office, for example, is currently developing measures of participation in voluntary and community-based activity.

If, at the end of the day, regeneration and local development are concerned with making areas more attractive to residents and to investors, it is these softer aspects that are at the heart of the long-term sustainable achievement of such change. As Solesbury (2002) suggests, the question has now changed from "what works" to "what works for whom, under what circumstances, and through which agencies". In tackling this, a blend of top-down quantitative and bottom-up qualitative evaluation methodologies is likely to prove appropriate.

References

AMION CONSULTING (2001), *Urban Regeneration Companies: Learning the lessons*, Department of the Environment, Transport and the Regions, London.

ARMSTRONG, H.W., KEHRER, B., WELLS, P. and WOOD, A.M. (2002), "The Evaluation of Community Economic Development Initiatives", *Urban Studies*, 39, 457-81.

ATKINSON, R. and KINTREA, K. (2001), "Disentangling Area Effects: Evidence from deprived and non-deprived neighbourhoods", *Urban Studies*, 38, 2277-98.

AUDIT COMMISSION (1989), *Urban Regeneration and Economic Development: The local government dimension*, HMSO, London.

AUDIT COMMISSION (2002), *Neighbourhood Renewal*, Audit Commission, London.

BEGG, I. (ed.) (2002), *Urban Competitiveness: Policies for dynamic cities*, Policy Press, Bristol.

BENNETT, T. (1996), "What's New in Evaluation Research? A note on the Pawson and Tilley article", *British Journal of Criminology*, 36, 567-73.

BRAMLEY, G., EVANS, M., ATKINS, J. *et al.* (1998), *Where Does Public Spending Go? Pilot study to analyse the flows of public expenditure into local areas*, Department of the Environment, Transport and the Regions, London.

CDP (1977), *Gilding the Ghetto: The state and the poverty experiments*, Community Development Project, London.

CURTIS, S., CAVE, B. and COUTTS, A. (2002), "Is Urban Regeneration Good for Health? Perceptions and theories of the health impact of urban change", *Environment and Planning, C*, 20, 517-34

DORLING, D., *et al.* (2001), "How Much Does Place Matter?", *Environment and Planning, A*, 33,1335-69.

DETR (1998), *Urban Development Corporations: performance and good practice*, Department of the Environment, Transport and the Regions, London.

DETR (2000), *Our Towns and Cities: The future. Delivering an urban renaissance*, CM4911, Department of the Environment, Transport and the Regions, London.

DTLR (2001), *Changing Fortunes: Geographic patterns of income deprivation in the late 1990s*, Department of the Environment, Transport and the Regions, London.

FALK, N. (2002), *Towns and Cities: Partners in urban renaissance*, URBED, London [*www.urbed.com*].

HALL, S. and NEVIN, B. (1999), "Continuity and Change: A review of English regeneration policy in the 1990s", *Regional Studies*, 33, 477-91.

HIGGINS, J., DEAKIN, J., EDWARDS, J. and WICKS, M. (1983), *Government and Urban Policy: Inside the policy-making process*, Blackwell, Oxford.

HM TREASURY (1995), *A Framework for the Evaluation of Regeneration Projects and Programmes*, HMSO, London.

HO, S.Y. (1999), "Monitoring and Evaluation of British Urban Policy: Case study of City Challenge", *Ph.D. thesis*, University of Manchester.

KPMG CONSULTING (1999), *City Challenge: Final national evaluation*, Department of the Environment, London.

LAWLESS, P., DABINETT, G., TYLER, P. and RHODES, J. (2000), *A Review of the Evidence Base for Regeneration Policy and Practice*, Department of the Environment, Transport and the Regions, London.

LENIHAN, H. (1999), "An Evaluation of a Regional Development Agency's Grants in Terms of Deadweight and Displacement", *Environment and Planning, C*, 17, 303-18.

MACLENNAN, D. (2000), *Changing Places, Engaging People*, York Publishing Services, York. [For details of the Joseph Rowntree publications from the Area Regeneration Programme, see *www.jrf.org.uk*.]

MCCULLOCH, A. (2001), "Ward-level Deprivation and Individual Social and Economic Outcomes in the British Household Panel", *Environment and Planning, A,* 33, 667-84.

MOORE, B. and RHODES, J. (1973), "Evaluating the Effects of British Regional Policy", *Economic Journal*, 83, 87-110.

NOBLE, M. *et al.* (2000), *Measuring Multiple Deprivation at the Small Area Level: The Indices of Deprivation 2000*, Department of the Environment, London.

PARKINSON, M.P. and ROBSON, B.T. (2000), *Urban Regeneration Companies: A process evaluation*, Department of the Environment, Transport and the Regions, London.

PAWSON, R. and TILLEY, N. (1994), "What Works in Evaluation Research?", *British Journal of Criminology*, 34, 291-306.

REGIONAL CO-ORDINATION UNIT (2002), *Review of Area Based Initiatives*, Regional Co-ordination Unit, Office of the Deputy Prime Minister, London.

RHODES, J., TYLER, P., BRENNAN, A., STEVENS, S., WARNOCK, C., OTERO-GARCIA, M. (2002), *Lessons and Evaluation Evidence from Ten Single Regeneration Budget Case Studies*, Department of the Environment, Transport, Local Government and the Regions, London.

ROBSON, B.T., BRADFORD, M.G. and DEAS, I. (1999), "Beyond the Boundaries: Vacancy chains and the Urban Development Corporations", *Environment and Planning* A, 31, 647-64.

ROBSON, B.T., BRADFORD, M.G, PARKINSON, M., *et al.* (1994), *Assessing the Impact of Urban Policy*, Department of the Environment, HMSO, London.

ROBSON, B.T., BRADFORD, M.G. and TOMLINSON, R. (1998), *Updating and Revising the Index of Local Deprivation*, Department of the Environment, Transport and the Regions, London.

ROBSON, B.T., BRADFORD, M.G. and TYE, R. (1995), "A Matrix of Deprivation in English Authorities, 1991" In *1991 Deprivation Index: A review of approaches and a matrix of results*, Department of the Environment, HMSO, London. pp.69-163.

ROSSI, P.H., FREEMAN, H.E. and LIPSEY, M.W. (1999), *Evaluation: A systematic approach*, 6th ed. Sage Publications, Thousand Oaks, California.

RUSSELL, H., DAWSON, J., GARSIDE, P. and PARKINSON, M. (1996), *City Challenge: Interim national evaluation,* Department of the Environment, London.

SOCIAL EXCLUSION UNIT (2001), *A New Commitment to Neighbourhood Renewal: National Strategy Action Plan*, Cabinet Office, London

SOLESBURY, W. (2002), "Evidence Based Policy: Whence it came and where it's going", *Planning Policy and Practice*.

STEWART, M., *et al*. (2002), *Collaboration and Co-Ordination in Area-Based Regeneration Initiatives*, Department of the Environment, Transport, Local Government and the Regions, London.

TYLER, P., RHODES, J. and BRENNAN, A. (1998), *The Distribution of SRB Challenge Fund Resources in Relation to Local Area Needs in England*, Department of the Environment, Transport and the Regions, London. [See also Brennan, A., Rhodes, J. and Tyler (1999) *Urban Studies*, 36, 2069-84.]

TYM, R. and PARTNERS (1984), *Monitoring Enterprise Zones: Three year report*, Department of the Environment, London.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How Assess What Works among Programmes and Policies
© OECD 2004

Chapter 9

A Commentary on Brian Rubson's Paper and the Workshop "Area-based Policy Evaluation"

by
Jonathan Potter,
OECD LEED Programme

This commentary focuses on three main points from Professor Robson's paper and the associated workshop discussion on the United Kingdom experience of area-based policy evaluation. They reflect issues that are also important for evaluation practice in other countries. Firstly, the quality and usefulness of area-based policy evaluation is often less than it could be because certain basic evaluation concepts are not being systematically addressed. Secondly, we should recognise that there exist a number of barriers to good quality area-based policy evaluation at local level, which limit the degree to which key evaluation concepts are employed. Thirdly, there are a number of ways in which local governments and development agencies can help to build evaluation capacity at local level, thus helping to overcome barriers to more rigorous area-based policy evaluation.

Key concepts for good quality area-based policy evaluation

As Professor Robson's paper points out, the past three decades have seen much area-based policy evaluation in the UK. Additionally, the arrival of the Labour government in 1997 brought a significant expansion in area-based policy evaluation work, associated with an expanded range of area-based programmes and a commitment to more evidence-based policy making across government as a whole. However, the evaluation that has been carried out has been of variable quality. Whilst Robson signals some major good practice evaluations, he also argues that many smaller scale evaluations have failed to grapple with certain conceptual issues that should be included in any rigorous area-based policy evaluation. Some key issues needing more attention in current evaluation practice are outlined below, drawing on and in parts extending the arguments in Robson's paper:

- *Programme rationale and objectives*. An evaluation should start by understanding the rationale and objectives originally set for a programme, *i.e.* what problems the programme was expected to address and how it was expected to address them. The underlying rationale may be articulated in policy documents but often it must be teased out by the evaluator from interviews with local policy makers. The evaluator should assess whether the rationale and objectives were appropriate and are still appropriate, and also compare programme results with the rationale and objectives to assess programme effectiveness in its own terms. A problem that often arises is that the rationale behind policy turns out to be misplaced to a greater or

lesser degree. Thus there has been much discussion of the tendency for policy makers to take solutions "off the shelf" or to follow fads, without properly thinking through whether the approach is really appropriate to the needs and opportunities of their area. A classic example is of localities seeking to attract high technology industry because of its associations with high wages and growth potential, even if the local conditions do not exist for sustainable development of the sector (university links, skilled labour, good communications, residential attractiveness, etc.). A related issue is consideration of alternatives, *i.e.* could there be alternative ways of achieving the same objectives more effectively? A rigorous evaluation should assess the programme rationale, considering whether it is appropriate and is the best way of meeting the objectives.

- *Deadweight*. Too many evaluations fail to consider the counterfactual, *i.e.* what would have happened in the absence of a programme. In particular, it is important to consider how far policy-supported actions would have taken place anyway (deadweight). Without such an assessment, it is easy to overestimate programme effects by attributing all of the positive outcomes in the area to the programme being evaluated. Where the counterfactual has been taken into account in UK evaluations, evaluators have tended to rely on self-assessments made by surveyed policy makers, programme managers, beneficiaries and other stakeholders. It is difficult to have full confidence in such evidence because respondents find it very difficult to answer hypothetical "what if" questions and because self-interest may lead to biased responses. Control group or control area comparisons can complement self-assessment exercises and provide the evaluator with greater confidence in results.
- *Displacement and substitution*. Displacement occurs where the start-up or growth of policy-supported enterprises/organisations leads to a loss of activity in other enterprises/organisations. In the context of labour market policy, substitution occurs where policy leads to supported individuals taking employment at the expense of non-supported individuals. Often these effects are not fully dealt with by evaluators. As with deadweight, evaluators often rely on self-assessment exercises about the degree to which other firms, individuals or areas are likely to have been adversely affected. Control group or control area comparisons should be used more often to complement this information. There has also been relatively little consideration to date of whether displacement and substitution should be discounted as entirely negative effects or whether they can be considered in part as a positive stimulus to dynamic adjustment and innovation in the local economy. The latter could be true if policy results in supported organisations and individuals developing competitive advantages over others, leading to higher productivity overall. Furthermore, it has been

argued that the mere creation of "churn" in the labour market could be beneficial if it protected individuals from the harm of being inactive for long periods of time. Thus the challenge for evaluators is firstly to measure the degree of displacement and substitution and secondly to decide how far this should be considered as a positive or negative phenomenon.

- *Longitudinal evaluations*. Area-based policy evaluations have often been limited to *ex post* assessments on completion of the programme. But it is difficult to track how well a programme has functioned if evaluation only takes place after the event, because of changes that take place during the programme, such as openings and closures of firms, movement of people and changes in the nature of supported activities. Some UK evaluations have attempted to develop a longitudinal approach, generally by making cross-sectional analyses of activity within the area at two or more points in time. For example, annual surveys of land use and enterprise activity on UK Urban Development Corporation zones helped to show the processes of change underway during the life of these initiatives. It is important to track programmes over time in this way.
- *Household mobility*. When examining how far policy is helping people living in marginalised neighbourhoods, evaluations must also confront the complex issue of the effects of household mobility into and out of the target area on evaluation results. The difficulty arises because assisted people may move out of the area and benefits accruing to them would not be assigned to the initiative unless the evaluation has some way of tracking the people assisted. Alternatively, a regeneration initiative may encourage people to move into the area, improving the population profile of the target area and giving the impression that policy has helped local people if incomers are relatively well off, although policy may not have directly helped the people originally targeted. In the UK, area-based policy evaluations have only rarely tried to track the movement of people that policy aims to help, but this should be included in the design of rigorous area-based policy evaluations.
- *Outcomes versus outputs*. Many evaluations stop at quantifying the outputs of a programme, for example the number of people obtaining a qualification, the number of people finding employment or the number of hectares of derelict land brought back into use. However, what is really of interest is the resulting outcome in terms of improving the quality of people's lives and stabilising or improving the economic and social vitality of the area. Assessing outcomes should therefore be the final target of area-based policy evaluation, concentrating on measures such as incomes, poverty, health, migration and land prices.

- *Disentangling causality*. Establishing causality is especially important for area-based policies because they tend to put in place multiple interventions addressing various aspects of economic, social and physical regeneration. The complication for evaluation is that actions aimed at one activity may have impacts on others. For example, measures for employment generation may reduce crime whilst measures that reduce crime may encourage new business activity. In order to judge the right policy mix for an area, policymakers needs to know the relative contribution of different actions to overall outputs and outcomes. There is still much progress to be made in measuring the extent and nature of synergies between actions. Sometimes matrix approaches are applied, indicating whether or not given actions contribute or are likely to contribute to improvements in more than one output field. This provides a framework for identifying possible synergies but does not go far in measuring their extent or how they occur. The need to measure synergies is increasingly recognised, but research is needed to develop more sophisticated techniques for doing so.
- *Small area data*. Because area-based policy tends to target small geographical areas whilst official data tends to be available only for larger administrative units, evaluators often face problems obtaining satisfactory data to measure the baseline situation and subsequent changes in the area compared with neighbours and the wider economy. As Robson notes, this issue has been recognised in the UK and considerable effort is going into improving the situation.
- *Scale of intervention*. Area-based policy evaluations tend to make their central reference the comparison of programme outputs or outcomes with programme costs. There are two main problems with this way of thinking. Firstly, there tends to be no assessment of whether the policy is being applied at sufficient scale to address the problems in a satisfactory timescale. A key question for evaluators should be, "if this policy is applied at this rate, how long would it take to resolve the problems of the area?" Secondly, there tends to be no assessment of whether, if resources were increased, an effective and efficient small-scale initiative could be scaled-up to deal with larger areas and client groups. Apart from the issue of whether any programme would work in the same way in a different context, the failure to consider scale can lead to misleading policy prescriptions suggesting that successful programmes naturally should be expanded. Small-scale schemes often appear very promising because they can deal with the areas and groups where policy is likely to make the most difference, but it might be difficult to achieve the same success with larger areas and groups.

Why is good quality area-based policy evaluation relatively rare?

Workshop participants sought to establish why good quality area-based policy evaluations, which take into account the sorts of conceptual issues referred to above, are relatively rare. The following possible barriers were identified:

1. Evaluation involves many positive externalities that are difficult for a local government or development agency to capture. For example, results are often made public, so it is easy for other areas to "free-ride" by referring to evaluations carried out by others.
2. There are economies of scale in evaluation that suggest they may best be undertaken at central level or by groups of local areas co-operating together rather than by individual local governments or agencies.
3. Good quality evaluation can be extremely expensive and can therefore become a significant proportion of total programme expenditure, especially for small programmes. In economic terms, this can be thought of as a problem of non-divisibility in that the costs of rigorous evaluation are difficult to reduce even for small programmes. Local agencies therefore have to balance the relative merits of expanding a programme or fully evaluating it. It is probably not wise to spend large amounts of money on evaluation of a very small programme.
4. In the United States, although less so in Europe and many other OECD countries, most economic development programmes are funded through forgone tax receipts rather than budget expenditure. The costs of economic development initiatives are thus less visible and often are not perceived as real costs by local officials and the electorate. There appears to be less pressure to evaluate programmes funded through forgone tax receipts, although pressure groups in the USA are beginning to draw attention to the issue.
5. Following a similar logic, often economic development carried out at the local level is not funded locally but from national, regional or other funds. Where little local money is being spent, local governments and agencies are less likely to wish to evaluate a local programme, although other funders may step in.
6. There is a fundamental mismatch between political and economic development timeframes. Thus, the political timeframe tends to be an election cycle, whereas economic development can take one to two decades. Local politicians are likely to be reluctant to sanction major long-term evaluations if they will provide little in the way of supporting evidence for their current activities.
7. Many local development professionals seem to be content to refer to case study examples of best practice rather than commissioning and using their own evaluations. This may reflect the lower cost of best practice exchange

and the attractiveness of keeping up with new trends presented in conferences and publications. But, whilst best practice exchange can be useful, especially when it is based on information from fully evaluated programmes, it is not a substitute for direct evaluations of local programmes. The problem here appears to be a scarcity of informed consumers to drive forward good quality area-based policy evaluation at local level.

Capacity building in area-based policy evaluation

In response to the above hypotheses, workshop participants suggested certain actions that might increase the quantity and quality of evaluation of area-based policies and their use by local policy makers and politicians in future programme design and implementation.

- Encourage mechanisms that lead to local funding of programmes, such as central government block grants that allow local governments and agencies flexibility in the choice and design of programmes for their areas or programmes that enable local governments to use retained locally-raised taxes for regeneration purposes. Local funding of programmes will create local pressure to evaluate that spending and lead to a greater sense of local ownership of the evaluation findings and commitment to using them.
- Encourage evaluations undertaken in collaboration by groups of local governments or development agencies or supported by a higher-level agency in order to address problems of externalities, economies of scale and high evaluation cost as a percentage of programme budgets.
- Address differences between the timescales of evaluators and politicians and policy makers to secure greater demand for and use of evaluation. This means stressing to politicians and policy makers that they should engage in economic development as a long-term exercise, for example by setting up arm's length development agencies with long-term goals, and that they cannot expect evaluators to demonstrate immediate success. At the same time, it should be stressed to evaluators that politicians and policy makers need interim results to help justify the continuation of programmes that appear to be working or to adjust programmes that appear not to be running well.
- Similarly, address differences of language between evaluators and politicians and policy makers, again to increase the demand for and use of evaluation, by increasing the understanding of politicians and their advising officials on evaluation techniques and implications and simplifying the way evaluators present their findings.
- Improve the capacities of local policy makers to commission and use area-based policy evaluations, including developing expertise on how to judge the quality of evaluation work and how to feed it into the next phase of

policy and practice improvement. Such client capacity building could work through education and information exchange to local development professionals. The following specific steps may be suggested:

- Forums could be developed in which experience could be exchanged between people involved in taking forward local area-based policies and using evaluations locally.
- Centres of excellence for training in area-based policy methods might be established, where the techniques and uses of evaluation could be an important component.
- Public dissemination of examples of good quality evaluations could be encouraged, for example using internet sites or magazines and publications aimed at local regeneration practitioners.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 10

Evaluating Business Assistance Programs

by
Eric Oldsman,
Nexus Associates Inc.
and
*Kris Hallerg**
Operations evaluation department, World Bank,
Washington, USA

* The views expressed here are those of the authors and do not represent the policies of the World Bank Group.

Foreword

Governments around the world are supporting a wide range of business assistance programs that aim to promote the development of private firms, particularly small and medium-sized enterprises (SMEs). Despite the level of resources committed to these programs, there has been relatively little effort devoted to determining whether these programs have indeed been successful in achieving intended outcomes. On the whole, evaluations have tended to rely on inherently flawed before-and-after studies or potentially biased testimonials from gratified customers.

However, there are better alternatives available to governments. Surveys of potential beneficiaries clearly have a place in evaluations, enabling evaluators to glean useful information on the perceptions of participants. But care needs to be given to ensure that surveys address critical aspects of program design, are worded in a way that takes the counterfactual directly into account, and are based on representative samples.

That said, under certain circumstances, participant judgment may not provide sufficient evidence of program impacts. Here, governments may want to employ more rigorous experimental or quasi-experimental designs to provide valid estimates of the impact of particular programs, controlling for extraneous factors that may influence observed changes in performance. Statistical techniques can be used to test various hypotheses concerning the impact of key variables. Moreover, to the extent possible, evaluations should include case studies based on rich narratives to explain causal mechanisms and identify elements of the program design that need to be modified. Finally, regardless of the particular approach, all evaluations should be based on clear statements detailing the target population, intended outcomes, and assumptions concerning the links between program activities and stated goals.

This paper seeks to provide government officials with a better understanding of the critical issues involved in program evaluation and the various tools that can be used in carrying out such studies. It focuses specifically on quantitative methods that can be used in summative evaluations of business assistance programs targeted to SMEs.

Introduction

Governments around the world have invested considerable amounts of money in a variety of initiatives to promote the development of private businesses, particularly small and medium-sized enterprises (SMEs).[1] The interest of federal, state and local governments in SMEs stems, in part, from the recognition that SMEs play a critical role in all economies – they produce a broad range of goods and services for domestic and foreign consumption, and in so doing, provide an important source of income and jobs in every region.

While SMEs constitute a significant share of the economy, many believe that the performance of SMEs is sub-optimal from a societal perspective and hold that government intervention is required to boost the growth and profitability of firms. Advocates for government intervention point to various imperfections in relevant markets as justifications for action.[2] In some cases, the argument is made that services needed by SMEs are not readily available in the market. In other cases, the rationale revolves around the contention that SMEs lack information required to make appropriate purchasing and/or investment decisions. In still others, the reasoning centres on the claim that decisions of individual private firms to pursue a particular course of action do not reflect broader social benefits or costs.

Decisions to fund initiatives targeted to SMEs are based on the belief that well-designed programs will address market imperfections, boost the performance of participating companies and yield significant economic and social benefits. Governments are now looking for credible evidence that these beliefs were right and that particular business assistance programs warrant continued support.

The call for good program evaluation reflects the fact that governments are constantly under pressure to allocate scarce resources to competing needs. These choices are rendered even more difficult, albeit necessary, in today's environment where discretionary spending is likely to be reduced significantly. Information on the actual results of programs established by government to meet various needs is critical to budget deliberations – evaluations can provide a basis for shifting resources away from under-performing programs to those that demonstrate success.

At the same time, governments need information to identify areas where changes in the program are required to improve the chances for success. Programs may need to be fine-tuned or subjected to substantial modifications based on hard evidence of what works, what doesn't and why. To this end, evaluation can help identify critical success factors and provide a basis for informed decisions on how best to redesign particular programs.

Decisions with respect to continued funding and/or ongoing operations should be based on accurate and credible information. In this regard,

evaluations need to be well-designed and implemented according to good standards of practice. As discussed in the next section, this should begin with a clear articulation of the program design.

Program design

SME initiatives are targeted toward firms (sole proprietorships, partnerships or corporations) within a specific size range as defined by annual sales, employment, and/or assets. Sometimes, other characteristics or conditions are used to define targeted SMEs. For instance, particular programs may target women-owned firms, rural enterprises, or specific industrial sectors.

Given some assessment of needs within the target population, governments have instituted programs that incorporate one or more of the following elements:

- Management or technical consulting services to address various business processes, including planning, product development, marketing and sales, production, distribution, human resources, information systems, and financial management.
- Training services to upgrade the skills of management, supervisors, machine operators, or other company personnel through some combination of classroom and hands-on instruction.
- Grants and other forms of concessional financing for capital investment, working capital requirements, or other needs.
- Tax credits for investment in research and development, capital equipment, and employee training.
- Access to low-cost facilities, equipment and other physical infrastructure.

All SME programs, regardless of which elements are included, are intended to yield desired outcomes at the level of the firm and broader economy. Specifically, these programs aim to change the behavior of participating firms, resulting in improved business performance and, in turn, improved economic and social conditions, as shown in Figure 10.1.

More detailed program logic models can be developed for specific programs. For example, as shown in Figure 10.2, an R&D tax credit targeted to all SMEs in a particular tax jurisdiction is intended to provide tax benefits to companies, thereby lowering the cost of R&D and inducing additional investment in product development efforts. This, in turn, is expected to lead to the actual development of novel products that meet customer needs and their subsequent introduction into commercial markets. It is further expected that these products will prove superior to other products in the market, generating increased sales and profitability for companies, as well as benefits for consumers. It is also assumed that companies that benefit from the R&D

Figure 10.1. **Basic program logic model**

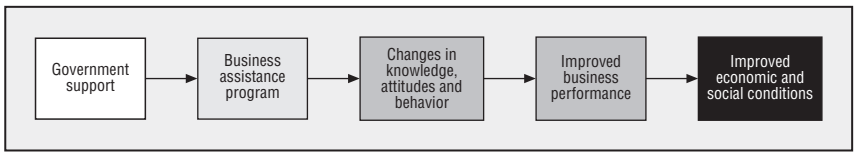
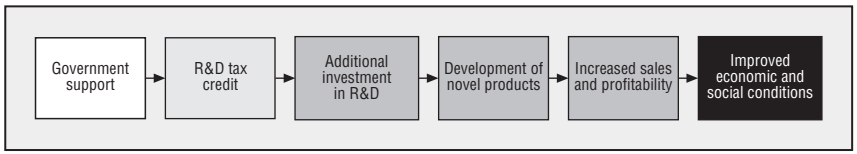


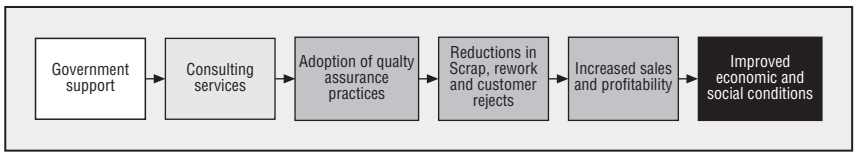
Figure 10.2. **Logic model for R&D tax credit for SMEs**



tax credit will hire additional researchers to undertake the expanded R&D program and other necessary employees to meet the growing demand for resulting commercial products. In this regard, it is assumed that products will be manufactured by companies directly benefiting from the R&D tax credit or licensees located in the same region or country.

In comparison, business consulting programs are intended to provide information to companies that lead the firms to effect changes in their operations that they otherwise would not have undertaken, yielding improvements in particular processes and overall enterprise-wide performance. For example, as illustrated in Figure 10.3, a program may centre on providing information on the importance of instituting sound quality assurance procedures. This is expected to lead companies to adopt particular practices such as statistical quality control or seek ISO 9000 certification. It is anticipated that the institution of these practices will result in improved quality as evidenced by reductions in the rate of scrap, rework and/or customer rejects. In turn, improved quality is expected to result in increased sales and profitability. Depending on the relationship between anticipated productivity gains and sales growth, programs may result in higher employment.

Figure 10.3. **Logic model for consulting program for SMEs**



These two examples demonstrate how program logic models provide concise descriptions of how programs will improve conditions within the target population, noting important causal mechanisms (If X, then Y). As a result, both examples present hypotheses that ostensibly could be tested in program evaluations. For example, in the case of the R&D tax credit, the key hypothesis is that the credit results in investments in R&D that otherwise would not have been undertaken by firms.[3] Similarly, for the consulting program, the principal hypothesis is that the service results in specific actions that otherwise would not have been undertaken by firms.

Therefore, before embarking on an evaluation, the target(s) of the program as well as the path linking activities to intended outcomes should be defined as clearly as possible.[4] The resulting program logic model should be used to define the scope of the evaluation, identify outcomes that should be measured, and help provide the basis for asserting causality.

Outcome measures

With a program logic model in hand, the next step is to establish a set of measurable indicators that can be used in assessing the impact of a particular business assistance program. In developing these measures, it is essential to consider the following:

- *Relevance*. Measures selected for the impact assessment need to be germane to the particular initiative being studied.
- *Validity*. Measures need to provide an accurate reflection of the underlying concept that is supposed to be measured.
- *Reliability*. Measures should be subject to as little measurement error as possible.
- *Practicality*. It has to be possible to obtain data needed to calculate measures.

The results of the evaluation will only be accurate and credible to the extent to which measures are relevant, valid, and reliable. But it also has to be feasible to employ measures given data availability, time, and budgetary constraints. For example, there are a variety of ways to measure productivity – *e.g.,* output per employee, value-added per labour hour, total factor productivity. The last measure reflects the additional value generated through the use of capital, labour, material and other factors of production. While it is arguably the best measure of productivity, it is very difficult to obtain required data even within large companies with sophisticated information systems. On the other hand, although output per employee as a measure of productivity may be misleading given that increased outsourcing will show up as a productivity gain, it is relatively simple to obtain necessary data. On balance, this may be the best choice for a specific impact assessment. Like other

aspects of designing and implementing evaluation, selecting outcome measures often involves tradeoffs.

Outcome measures need to be developed within the context of particular initiatives, reflecting the specific targets and goals of the intervention as well as practical concerns with respect to data availability. There is no one set of measures that will fit all business assistance programs initiatives. But, there may be similar indicators for similar programs. Examples are shown in Table 10.1. Many of these indicators focus on changes in quality, turnaround time, and production costs. Others are intended to measure enterprise-wide performance with respect to changes in sales, net profits and employment.

Table 10.1. **Potential outcome measures for targeted SMEs**

Indicator	Definition
Attitudinal changes	Prevalence and incidence of particular attitudes among managers, supervisors and/or workers.
Process changes	Prevalence and incidence of changes in particular processes, *e.g.* planning, sales and marketing, production, and distribution.
Investment	Dollars invested in plant, equipment, software and/or training.
Defect rate (rework or scrap)	Proportion of units that do not conform to design standards and are subsequently reworked or scrapped.
Order-to-delivery time	Total amount of time (hours or days) from receipt of order to delivery at customers' premises.
On-time delivery rate	Proportion of orders delivered to customer according to agreed schedule.
Customer rejects	Proportion of items delivered to customers and subsequently rejected due to nonconformity.
Capacity utilisation	Proportion of available resources (*e.g.* plant and equipment) used in production.
Labour productivity	Sales value of output produced during the period divided by direct labour hours used in production.
Sales	Revenues derived from the sale of goods or services.
Net profit	Operating profit (sales minus cost of goods sold) and other income less total expenses.
Employment	Full- and part-time workers employed by companies or sole proprietorships as of a specific date or pay period, *e.g.* the week of March 12th.

Methods for assessing impacts

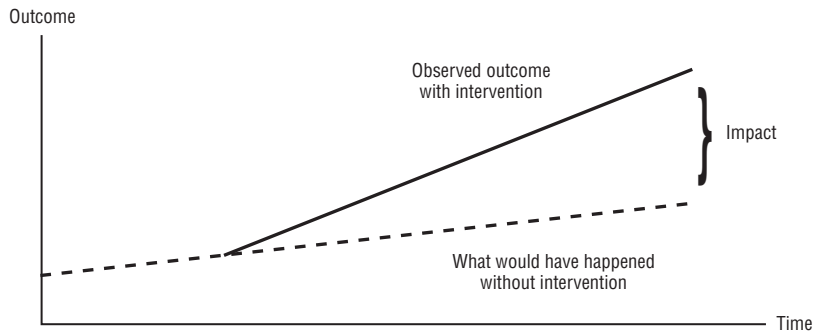
Impact assessments are undertaken to find out whether a program actually produced intended outcomes. In demonstrating that a particular intervention resulted in a specific outcome, certain conditions need to be met.[5] First, changes engendered through the intervention have to be shown to produce the effect – put another way, the outcome must be responsive to the intervention. Second, plausible alternative explanations for the observed outcome have to be ruled out – rival hypotheses must be disproved. Third, the mechanism by which the outcome was produced has to be explained – in

other words, a theory linking the intervention to the outcome must be articulated. Finally, it must be possible to replicate the results in similar settings. With proper research, apparent correlations can be translated into credible causal explanations.

In this regard, the fundamental tenet of impact assessment is the need to compare the observed situation with the intervention to what would have been had there been no intervention at all. The difference in resulting outcomes between these two states constitutes the impact of the intervention as illustrated in Figure 10.4.[6] While the counterfactual cannot be observed or known with complete certainty, the concept of comparing observed outcomes to this hypothetical state underlies all valid approaches to assessing impacts. Valid comparisons imply that the net effect of interventions is isolated from all other extraneous or confounding factors that influence defined outcomes. For example, efforts to improve the performance of firms by providing vouchers for consulting services may have been undertaken during a time of rapid economic expansion buoyed by substantial tax breaks, aggressive regulatory reform, and booming consumer demand. Given these conditions, it is likely that participating firms would have enjoyed significant growth even in the absence of the voucher program. As a result, the central question is not whether participating firms grew, but rather did these same firms grow more than would have been expected if they had elected not to participate in the voucher program. Thus, the major challenge in impact assessments is to estimate the effect of programs after netting out extraneous factors that affect outcomes. These factors may include specific events or long-term trends in particular industries, regions or countries as in the example cited above. They may also include ongoing developments within participating SMEs.[7]

Similarly, to the extent possible, impact assessments need to account for the voluntary nature of programs. SMEs take part in programs of their own

Figure 10.4. **The impact of an intervention**



volition. Some members of the target population may be more inclined to participate due to greater interest, motivation or other conditions within the firm. This self-selection process can bias results if the factors that lead companies to participate are related to the specific outcomes under study.[8] For example, initiatives that focus on providing greater access to long-term financing for the purchase of fixed assets are likely to attract growing companies with progressive management that recognise potential market opportunities, are willing to assume certain risks in the hope of reaping financial returns, and have sufficient collateral to secure the loan. These same characteristics are likely to be associated with future sales growth. It would be inappropriate to compare this segment of the population of firms to other SMEs that may be struggling to survive. To do so would run the risk of overestimating the impact of the financial assistance program. As discussed below, care needs to be taken to account for potential selection bias in estimating the impact of business assistance programs.

While there are numerous variations, the menu of options available to assess initiatives targeted to SMEs is limited to four basic methods based on the type of controls used to isolate program effects from other confounding factors – experiments with random assignment, quasi-experiments with constructed controls, participant judgment and expert opinion, and non-experiments with reflexive controls.[9] The strength of causal inferences that can be drawn from the analysis depends on how well the particular approach used in assessing impacts deals with the threats to validity.[10]

Regardless of the purpose or design of the initiative, all impact assessments need to employ one or more of the following methods:[11]

1. *Experiments with random assignment*. The gold standard in impact assessment is experimental design with random assignment to treatment and control groups. In this approach, SMEs in the treatment group receive assistance; those in the control group receive an alternative type of assistance or none at all. The critical element of this design is randomisation. Random in this case does not mean haphazard; care needs to be taken to ensure that every company has an equal chance of being selected for either group. Random assignment helps guarantee that the two groups are similar in aggregate, and that any extraneous factors that influence outcomes are present in both groups. For example, random assignment helps ensure that both groups of SMEs are similar in terms of the proportion of firms that are inherently more receptive to making needed changes in business practices, or that fluctuations in market conditions affect both groups equally. As such, the control group serves as the ideal counterfactual. Because of this comparability, claims that observed differences in outcomes between the two groups are the direct result of the program are more difficult to refute.

Evaluations using experimental designs are quite common in the health, social welfare and educational arenas to test the efficacy of new approaches (see Box 10.1) However, although this approach is very strong, it has not been used extensively in evaluating the impact of business assistance programs. There are several reasons for this. First, political considerations sometimes make it difficult to assign SMEs to different groups: politicians and program managers are hesitant to provide different services or deny service altogether to companies randomly assigned to the control group. Second, it is frequently hard to maintain experimental conditions: although SMEs may be statistically equivalent at the start of the program, some participants may refuse to participate or may drop out of the program. Moreover, the services provided to SMEs may not be standardised and may change over time as programs evolve. Finally, evaluations using experimental design tend to be costly and difficult to administer.

Box 10.1. **Examples of experimental designs**

Argentina workfare-to-work experiment.* The Proempleo program provided a wage subsidy and specialised training as a means of assisting the transition from workfare to regular work. Participants were located in two adjacent municipalities and were registered in workfare programs. Workfare participants (958 households) were randomly assigned to one of three roughly equal-size groups: *a*) those that were given a voucher that entitled an employer to receive a sizable wage subsidy, *b*) those that received voluntary skill training along with the voucher, and *c*) those that received no services and served as the control group.

The evaluation attempted to measure the direct impact of the experiment on the employment and incomes of those who received the voucher and training. A baseline survey and several follow-up surveys were conducted over 18 months. Double-difference and instrumental-variables methods were used to deal with potential experimental biases, including selective compliance. Compared to the control group, voucher recipients had a significantly higher probability of employment after 18 months, though their current incomes were no higher. The impact was largely confined to women and younger workers.

* Galasso, Ravallion, and Salvia (2001).

2. *Quasi-experiments with constructed controls*. In situations where experimental design is infeasible or impractical, the next preferred approach is a quasi-experimental design. As in the previous design, the change in the performance of participating SMEs is compared to other similar SMEs that

have not received assistance. However, in this case, assignment to the two groups is non-random. Rather, a comparison group is constructed after the fact. To the extent that the two groups are similar, observed differences can be attributed to the program with a high degree of confidence. Valid comparisons require that the two groups be similar in terms of their composition with respect to key characteristics, exposure to external events and trends, and propensity for program participation.[12]

There are several types of designs that fall within this general category. These are discussed below in the order of their ability to deal with confounding factors:

- *Regression discontinuity*. In this approach, scores on a specific measure are used to assign targets to the intervention and control groups in an explicit and consistent manner. The difference in post-implementation performance between the two groups is compared, statistically controlling for the variable used in the selection process. For example, scores with respect to the creditworthiness of SMEs may be used to qualify firms for participation in a loan assistance program – a case of administrative selection. Assuming that an explicit cut-off point is used to determine eligibility, the net effect of the program can be estimated after adjusting for the original selection variable.
- *Statistically equated controls*. This approach employs statistical techniques to ensure that the intervention and control are as equivalent as possible with respect to outcome-related characteristics. In general, this involves using multivariate regression in which the influence of the program is estimated after controlling for other variables that may affect outcomes. For example, the statistical model used to estimate the effect of a consulting program on firm productivity may include various control variables such as firm size, industry classification, geographical location, ownership, and initial capital stock, as well as factors influencing selection. Selection is addressed through the use of two-stage regression or other techniques involving instrumental variables.[13] In the two-stage approach, an initial equation is used to model the selection process. The result of this analysis (inverse Mills ratio) is then incorporated into a second equation along with other control variables to estimate outcomes. As such, this approach explicitly accounts for potential selection bias.
- *Matched controls*. A somewhat less sophisticated approach involves constructing a comparison group that resembles the treatment group as closely as possible based on characteristics considered important in explaining outcomes. For example, companies may be matched based on the same set of variables described in the previous technique. Performance differences between the two groups post-intervention are calculated without

further statistical adjustment. However, it can be difficult to find matches for participants that are simultaneously based on all criteria, *e.g.,* another company of the same size, industry, geographical location, ownership, etc.

- *Generic controls*. The last approach uses measurements of performance for the population from which targets are drawn as a control. For example, annual sales growth among participating enterprises may be compared to industry averages, with any resulting difference attributed to the program. However, generic controls may not be capable of ensuring comparability with participants and should be used with caution.

Despite their complexity, quasi-experimental designs have been used in evaluating a broad range of development assistance programs. Examples are shown in Box 10.2.

- *Participant judgment and expert opinion*. The final approach relies on people who are familiar with the intervention to make judgments concerning its impact. This can involve program participants or independent experts. In either case, individuals are asked to estimate the extent to which performance was enhanced as a result of the program – in effect, to compare their current performance to what would have happened in the absence of the program.

  While this approach is quite common, it is fraught with problems. It requires people to be able to determine the *net* effect of the intervention based solely on their own knowledge without reference to explicit comparisons. However, it may be the only option available given data and budget constraints. When used, care should be taken to make sure that people consider the counterfactual in their assessment of impacts (see Box 10.3).

- *Non-experiments with reflexive controls*. Before-and-after comparisons are generally invalid because they fail to control for other factors that may have contributed to observed outcomes. As such, results from studies based exclusively on reflexive controls should be treated with substantial skepticism. That said, this approach may be valid when there is a clear and close relationship between the program and outcomes of interest (see Box 10.4). In addition, reflexive controls are sometime used when it is impossible to construct a control group as is the case for full-coverage programs that affect all companies in the target population.

In all four approaches, it is possible to use program data to enhance the analysis. It is often the case that programs are not administered uniformly – that is, the intervention may vary in intensity across members of the target population. For example, while some SMEs may receive 40 hours of technical assistance under a scheme to provide consulting services on a cost-shared basis, others may receive significantly more or less assistance. The impact of varying levels of intensity (sometimes referred to as the dosage effect) can be

Box 10.2. **Examples of quasi-experimental designs**

Industrial Resource Center (IRC) program.[1] The program was established in 1988 to help small and medium-sized manufacturers upgrade business practices and modernise their production capabilities in order to spur economic growth in Pennsylvania. The IRCs are designed to accomplish this mission through a comprehensive set of activities involving a combination of consulting and training services. Since its inception, the state government has committed roughly $84 million to the program.

A comprehensive evaluation of the program was conducted in 1999. The evaluation included an assessment of the impact of services on participating companies with respect to growth in productivity and output. To help control for potential selection bias, the analysis employs a two-stage procedure.[2] The first step involves estimating the probability of companies becoming IRC clients as a function of characteristics of the firm. The second step involves estimating the impact of the IRC program on companies after controlling for factors that affect productivity and output growth as well as potential selection bias. The estimated model is based upon a modified Cobb-Douglas production function which includes plant specific factors.

The analysis is based on panel data for individual manufacturing plants – the Longitudinal Research Dataset (LRD) – maintained by the Center for Economic Studies at the US Bureau of the Census. The dataset provides detailed plant-level data on shipments, employment, factor costs, industry, and other legal and administrative identifiers. It is compiled from the Census of Manufactures carried out every five years and the Annual Survey of Manufactures. The LRD was used to obtain data for both clients and non-clients in Pennsylvania. Companies included in the IRC administrative database were linked to the LRD using a matching procedure developed by the Center for Economic Studies. The matching process identified 2 839 unique IRC client establishments in the census years based on its permanent plant number (PPN). The comparison group included a similar number of companies.

The analysis demonstrated that the program had a significant impact on IRC clients. For the pre-92 cohort, the difference in the growth rate in output and productivity directly attributed to IRC services is estimated at 1.8 per cent and 3.6 per cent per year over a ten-year period.

Small Business Innovation Research (SBIR) program.[3] The SBIR program was established by US Congress in 1982. The authorizing legislation mandated that all federal agencies spending more than $100 million annually on external research set aside a fixed percentage of these funds for awards to small businesses. Over time, the percentage has increased to 2.5 per cent. Between 1983 and 1995, small firms received more than $6 billion under the program.

Box 10.2. **Examples of quasi-experimental designs** *(cont.)*

An evaluation of the program was conducted in 1996, focusing on the impact of SBIR funding on sales and employment growth. The analysis involved comparing the performance of firms that received SBIR grants to similar companies that did not participate in the program, controlling for firm age, geographical location, prior venture financing, and overall venture capital activity in the region and industry. The sample consisted of 541 firms that received SBIR Phase II awards in the first three program cycles and 594 matching firms.

Data was compiled from program records and information contained in publicly available directories and databases.

The analysis demonstrated that SBIR awardees experienced greater growth in both sales and employment than similar firms, but these effects were confined to areas that attracted significant venture financing.

1. Nexus Associates, Inc. (1999).
2. See the following publications for additional information on procedures used in this analysis: Maddala (1994), Jarmin (1997), and Heckman (1974 and 1979).
3. Lerner, Josh (1996).

examined under all four approaches. Depending on the degree of variation, this approach can strengthen the causal inferences that can be drawn from the analysis.[17]

Selection of an appropriate method

The four approaches to assessing impacts can be applied to a variety of questions that might be posed by governments with respect to business assistance programs. The choice of the approach to use in a particular impact assessment needs to take several factors into account. Each approach has strengths and weaknesses as illustrated in Figure 10.5. For example, experimental designs provide strong evidence of causality, but may be expensive and difficult to administer. Non-experimental designs are generally easier to implement, but may not offer strong enough causal inferences.

The figure suggests that more sophisticated approaches such as experimental and quasi-experimental designs should be used wherever warranted given the strength of the causal inferences that can be drawn. However, the additional strength comes at a higher cost. Therefore, these approaches would be appropriate only when further significant investments are being considered. For example, numerous initiatives are established initially as pilot programs with the expressed intention of expanding and/or replicating the initiative if successful. Depending on the magnitude of the

Box 10.3. **Examples of participant judgement and expert opinion**

Regional Enterprise Grants.[1] The grant program was initiated in 1988 by the Department of Industry and Trade in the UK. The program has two components: a flat rate grant contributing 15 per cent (up to a maximum of £15 000) toward the purchase of fixed assets, and grants covering 50 per cent of eligible project costs (up to a maximum of £25 000) to develop new product or processes to the point of commercial production. Firms employing fewer that 25 are eligible for participation.

An evaluation of the program was conducted in 1991. It was based on the results of surveys of a representative sample of 100 firms drawn from the population of companies that had submitted applications after 31 December 1998 and completed projects. All of the surveys were conducted in person. The survey asked respondents to indicate the impact of the grants in terms of increased turnover, value added, and employment. The counterfactual was addressed through in-depth questioning about what firms would have done if no grant had been available and how they would have financed their projects in the absence of the grants.

The study suggested that the impact of grants for fixed assets was fairly weak, whereas grants for innovation have performed a useful role among firms with genuine growth potential.

Manufacturing Extension Partnership. In 1988, US Congress directed the National Institute of Standards and Technology (NIST) to establish a program to help small manufacturers – manufacturing establishments with 500 or fewer workers – to improve their competitive performance. Beginning with just three centres, the NIST Manufacturing Extension Partnership (MEP) has expanded to 68 centres with over 400 offices throughout the United States.

All companies that received services from a NIST MEP centre are surveyed by an independent survey firm one year after the completion of a major project. Major projects are defined as activities requiring eight or more hours of centre staff or third-party service provider time. In each quarterly survey, clients are asked about impacts experienced in the previous twelve months as a direct result of services provided by the centres. Specifically, they are asked to estimate changes in performance resulting from participation in the program. Survey results suggest that aggregate impacts with respect to increased sales and employment are quite large relative to the federal investment in the program: however, a small share of firms that responded to the survey account for the bulk of aggregate impacts.

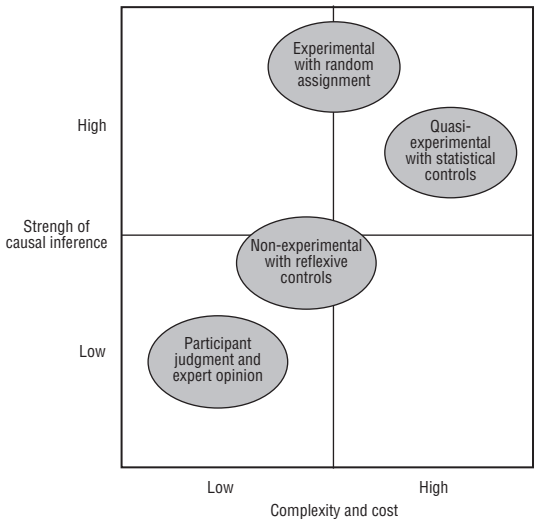
In addition to quantitative techniques, NIST MEP has supported case study research focusing on exemplary engagements based on a conceptual model linking services to program outcomes.[2]

1. Segal Quince Wicksteed, Ltd. (1991).
2. US Department of Commerce (1997).

Box 10.4. **Example of non-experiments with reflexive controls**

COMPITE. In the COMPITE program in Mexico, certified consultants work with companies over a four day period to improve the performance of one production line along four dimensions – productivity, manufacturing lead time, work-in-process inventory, and floor space requirements. The first day is spent training up to 20 employees in lean production techniques. The second day is spent measuring current performance, diagnosing problems, and devising means to improve performance. Day three is devoted to implementing changes. Another set of measurements are taken on the fourth day and compared to those taken before the changes were made. While this is a before-after design, because of the proximity of the program intervention there is little doubt that the program caused the changes in observed outcomes. The before-after design would not be particularly strong, however, if the question turned to the impact of COMPITE on enterprise-wide outcomes one or two years after the company participated in the program.

Figure 10.5. **Trade-offs in evaluation design**



required investment, it may make sense for governments to commit resources for a rigorous evaluation to help inform decisions. However, the ability to use these approaches is contingent on whether requisite data can be obtained and whether there is sufficient time to design and implement the study.

The general rule should be to use the best possible design from a methodological perspective, taking into account the significance of the investment, as well as practical considerations related to technical feasibility.

Data collection

Regardless of the approach, all impact assessments require accurate and reliable data. At a minimum, it is necessary to collect data on outcome measures for entities that were affected by the initiative. If comparative analyses are contemplated, data on outcome measures and other variables will be needed for members of the treatment and control groups.[18, 19] In all cases, baseline data (preferably multiple measures) are needed along with data after the intervention.

There are only two options for obtaining requisite data:

1. *Draw on existing data maintained by government and other organisations*. Given the cost of data collection, it is preferable to take advantage of existing data to the extent possible. Government statistical agencies in many countries conduct surveys of enterprises and households on a routine basis that might be used in impact assessments. These include national income and expenditure surveys, household income and expenditure surveys, labour market surveys, and various industrial surveys. Other organisations such as banks, credit unions and cooperatives maintain data on large numbers of companies as part of ongoing operations. These sources should be explored to see whether data required for the impact assessment are available.
2. *Commission special surveys*. In many cases, however, the only recourse will be to conduct a survey undertaken specifically for the impact assessment. There are a number of critical issues that need to be addressed to design and administer a survey successfully. The type of survey selected, the wording of questions, sampling strategies, follow-up, and data entry procedures all have an important bearing on the accuracy and utility of survey results. A special word on sampling is also in order. It is essential to use probability sampling to ensure valid results; stratification should be considered for greater efficiency and to ensure that the sample accurately represents the overall target population.[20]

In addition, all four approaches require administrative records to identify and characterise service recipients as well as the nature of services received. Moreover, certain techniques require additional information. For example, in order to employ regression discontinuity, the program must maintain data on variables used to determine eligibility and/or qualification for participation.

Recommendations

The basic principals and techniques for conducting evaluations are well established. However, only recently have they been applied to business

assistance programs. This experience points to several important recommendations:

- *Clarify targets, goals, and underlying program logic*. Impact assessments require a careful articulation of the targets of the initiative, the specific changes that are expected to be brought about as a result of the initiative, and the causal relationships between particular activities and intended outcomes. This should be summarised in a formal program logic model or log frame. This exercise is best done as part of the process of designing new initiatives, rather than as the initial task of an *ex post* evaluation.
- *Plan evaluation at the inception of programs*. Impact assessment should be planned as early as possible, preferably before the initiative has been launched. This is clearly the case for experimental approaches with random assignment; however, in general *ex ante* designs tend to be stronger since measures for collecting required data can be put in place prior to program implementation. To this end, all programs should be required to develop a formal evaluation plan as a condition for funding. The plan should discuss the purposes of the evaluation; specific questions that will be addressed; evaluation design; process and outcome measures; data collection strategy; possible analyses; reports and other methods of communicating results; timeline; roles and responsibilities of staff and outside contractors, if any; and an estimated budget.[21]
- *Establish baseline data and program records*. Programs should collect baseline data on characteristics and performance of program targets. Moreover, all programs need to maintain complete and accurate records as part of program implementation, including the nature and magnitude of resources committed to particular companies and/or institutions.
- *Recognise that impact assessment is explicitly about demonstrating causality*. While some people within the donor community (funding development co-operation in developing countries) are calling for greater accountability in terms of effectiveness, others bemoan the futility of trying to establish attribution. It is difficult to reconcile these views. Impact assessments are concerned specifically with demonstrating that particular initiatives produced the desired results – put another way, they aim to establish causality. All evaluations should be designed to show effects, rule out alternative explanations, and explain causal mechanisms. The replication of results in similar settings can add further credibility.
- *Build valid comparisons into the analysis*. Assessing the impact of initiatives targeted to SMEs involves comparing observed phenomenon to the counterfactual – a hypothetical situation that would have occurred in the absence of the program. Random assignment, constructed controls, and/or

reflexive controls are needed to isolate the impact of the program from other factors affecting outcomes.

- *Use multiple methods*. The approaches described in the paper are not mutually exclusive. Wherever possible, multiple techniques should be used to assess the impact of particular initiatives. Similar results from different methods can add to the credibility of findings. Moreover, qualitative research should be used to complement quantitative techniques, providing insights into the specific causal mechanisms that come into play in generating outcomes.
- *Recognise that good enough is good enough*. Governments and other stakeholders should strive for as much rigor as possible. However, practical considerations need to be taken into account in designing and implementing an impact assessment. Data, time, and budgetary constraints may make it infeasible to adopt certain approaches. Stakeholders need to accept these limitations.
- *Commit resources*. The amount of money devoted to evaluations is at the discretion of governments. Governments must be willing to commit the level of resources needed to design strong evaluations, collect vital data, conduct required analyses, and report results.

Notes

1. Definitions vary by country and are usually based on the number of employees, annual sales, or assets. Typically, microenterprises are defined as firms with up to 10 employees, small enterprises have from 10 to 50-100 employees, and medium enterprises have from 50 to 100-250 employees.
2. Economics teaches that perfect competition will lead to Pareto-efficient allocation of resources as long as certain assumptions are met. It requires that a large number of producers and consumers exist in a given market, none of which can influence price on their own; economic actors are rational; all resources (including information) are perfectly mobile; and transaction costs are zero. Unfortunately, markets in the real world are never in accord with the ideal. Moreover, Pareto-efficiency does not guarantee an equitable outcome. Three types of market imperfections are relevant to business assistance programs. Information asymmetries may lead companies to forego needed assistance or investments. In addition, companies may not be able to appropriate the full benefits associated with a particular action, leading to underinvestment from a societal perspective. These externalities arise when the production (or consumption) of a good affects parties other than those directly involved in the transaction. Finally, in some cases, business assistance may constitute a public good. Public goods have two unique properties: First, consumption of a public good by one consumer does not affect the ability of other consumers to benefit from it (*non-rivalry*); and second, it is difficult to stop people from benefiting from the good even if they are unwilling to pay for it (*non-excludability*). Because of these two characteristics, public goods tend to be undersupplied in a market economy.
3. More formally, evaluations should be designed to test the null hypothesis that the R&D tax credit does not lead to any additional investment. Qualitative and/or

quantitative techniques are used to determine whether the null hypothesis can be rejected at a reasonable level of confidence.

4. Preferably, a program logic model should be developed during the project design stage.

5. See Mosteller and Tukey (1977) for a discussion of conditions required to demonstrate causality.

6. This is sometimes referred to as "additionality".

7. Threats to internal validity are generally grouped under several broad categories such as external events, secular drift, maturation, regression and attrition. Readers interested in exploring these concepts in more detail are referred to Cook and Campbell (1979).

8. A similar sort of selection bias can occur when organisations select participants based on certain characteristics – this is referred to in the literature as administrative selection.

9. A fifth approach — structured case studies – can also be used to examine the impacts of intervention on participants and can be used to supplement quantitative techniques. Unlike the other approaches described in the paper, case studies rely on extensive narrative descriptions and other evidence to assert that the intervention caused observed outcomes. In general, case studies involve multiple sources of information including direct observation, interviews, documents, and physical artifacts. In all instances, program logic models play a critical role. While case studies can provide rich explanations of how and why the program affected particular firms, it is difficult to generalise results beyond the firms studied. This is particularly true for programs with diverse clients and services.

10. It is important to note that these approaches are not mutually exclusive; they can be combined under certain circumstances to enhance the analysis.

11. Some of the methods to assess the impact of business assistance programs are quite complex, requiring a background in statistics and econometrics. As such, a detailed explanation of the technical issues involved in carrying out each type of study is outside the scope of this paper.

12. The issue of the validity of a comparison group is central to this approach. Ideally, the non-participant group should be similar to the participant group with respect to variables affecting outcome measures, but should not have received business assistance through the government program. "Similar" in this context refers to the distribution of values for these variables, *i.e.*, the mean and range.

13. See Heckman (1985). A similar technique known as Propensity Score Matching can also be used to control for selection. See Jalan and Ravallion (forthcoming).

14. Nexus Associates, Inc. (1999).

15. See the following publications for additional information on procedures used in this analysis: Maddala (1994), Jarmin (1997), and Heckman (1974 and 1979).

16. Lerner, Josh (1996).

17. The use of dosage data is particularly important in the case of full-coverage programs where pure comparison groups are unavailable.

18. Data requirements are specific to the outcome measures of interest and the nature of the analysis that will be conducted. In addition to outcome measures,

data may be required for explanatory variables used in regression analyses, including instrumental variables used to control for potential selection bias.

19. Large samples are generally required for quasi-experimental designs. In computing the required sample size, it is important to consider three factors – the likely variance in the outcome measure, the required confidence level, and the desired precision of the estimate. The latter can be thought of as the minimum effect desired by stakeholders – anything less would not be considered successful or of particular interest. A concrete example: assume that stakeholders are interested in determining whether the program has resulted in increased sales growth. Further, assume that the program should aim to increase growth by 5 percentage points more than would have occurred in the absence of the program. Put another way, if companies in the comparison group grew by an average of 10 per cent per year, stakeholders would like to see participating companies grow by an average of 15 per cent annually. Given a standard deviation of 30 per cent, a 95 per cent confidence interval and a test power of 90 per cent, data would need to be obtained from roughly 620 participating and non-participating companies.

20. Kish (1965).

21. The evaluation plan should go beyond issues related to impact assessments and address other facets of performance, including outreach, operating efficiency, financial self-sufficiency and other issues.

References

COOK, Thomas D. and Donald T. CAMPBELL (1979), *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

GALASSO, Emanuela, Martin RAVALLION and Agustin SALVIA (2001), "Assisting the Transition from Workfare to Work: Argentina's Proempleo Experiment". World Bank, September 24.

HECKMAN, J. (1974), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and Simple Estimator for such Models", *The Annals of Economic and Social Measurement*, 5: 475-492.

HECKMAN, J. (1979), "Sample Selection Bias as Specification Errors". *Econometrica*, 47: 153-161

HECKMAN, James and Richard ROBB (1985), "Alternative Methods of Evaluating the Impact of Interventions: An overview". *Journal of Econometrics* 30: 239-67.

KISH, Leslie (1965), *Survey Sampling*. New York: John Wiley and Sons.

JALAN, Jyotsna and Martin RAVALLION (forthcoming), "Estimating the Benefit Incidence of an Antipoverty Program by Propensity Score Matching", *Journal of Business and Economic Statistics*.

JARMIN, R.J. (1997), "Evaluating Impact of Manufacturing Extension on Productivity Growth", Manuscript, Center for Economic Studies, US Bureau of Census, Washington DC. 1997.

LERNER, Josh (1996), "The Government as Venture Capitalist: The long-run impact of the SBIR Program", Working Paper 5753, National Bureau of Economic Research.

MADDALA, G.S. (1994), "Limited Dependent and Qualitative Variables in Econometrics", *Econometric Society Monographs* No. 3, Cambridge University Press. 1994.

MOSTELLER, F.M. and J.W. TUKEY (1977), *Data Analysis and Regression*. Reading, MA: Addison-Wesley.

NEXUS ASSOCIATES, INC. (1999), "The Pennsylvania Industrial Resource Centers: Assessing the record and charting the future".

SEGAL QUINCE WICKSTEED, LTD. (1991), "Evaluation of Regional Enterprise Grants – Second Stage".

US DEPARTMENT OF COMMERCE (1997), *MEP Successes: A case study approach*. Gaithersburg, MD: NIST Special Publication 916.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 11

Evaluating Training Programs: Impacts at the Local Level

by
Randall W. Eberts
and
Christopher J. O'Leary
W.E. Upjohn Institute for Employment Research, Kalamazoo, MI, USA

Introduction

Evaluations are an integral part of many of the public job training programs in the United States. Since the 1960s, considerable time and resources have been devoted to better understanding the impacts of employment and training programs. Indeed, many advances in techniques for evaluating social programs were developed in the course of studies into employment program effects (Stromsdorfer and Farkas, 1980). Policy makers came to appreciate the value of objective and rigorous evaluations in helping to guide their decisions regarding these programs. In fact the legislation authorizing the Job Training Partnership Act (JTPA) of 1982 mandated an ongoing performance measurement system and a net impact evaluation. The latter was to be done using the method which emerged as the gold standard for evaluation – a field experiment involving random assignment (Orr *et al.*, 1996). As noted by Kluve and Schmidt (2002), the American emphasis on evaluating employment and training programs is distinctly different from the usual practice in Europe.

However, even in the United States enthusiasm for evaluations does not extend to all job training programs. While most federally funded programs are regularly evaluated, state supported programs have received far less scrutiny. This is particularly true for those state sponsored training programs that provide customized training to firms as a way to promote local and regional economic development. One reason may be the added complexity of evaluating these programs which commonly have a multiplicity of objectives. Evaluation of such program designs often result in imprecise and inconclusive results, which undermines the credibility and value of doing evaluations. No less important may be the lack of political will and limited resources to conduct appropriate evaluations at the state and local level.

This paper reviews previous evaluations of job training programs with an eye to lessons that can be applied to state and local programs. In a real sense, all evaluations are local, because participation takes place and data are collected locally. Evaluations of federal job training programs typically focus on the average success of program participants, while state and locally-financed programs often seek to affect the overall condition of a local economy. We contrast evaluations of federal job training programs with those of state-financed training programs to highlight how evaluation techniques vary to accommodate different features of programs and the context in which they are administered.

In evaluating job training programs, or any government program for that matter, several basic steps must be taken, and we organize the paper around these components of an evaluation process. First, one must understand the purpose of the program so that appropriate outcome measures can be quantified and the proper data collected. Second, the administration of the program must be understood, including the way in which services are delivered. Third, one must also identify the groups that are targeted for services and the characteristics of those who actually participate. Finally, the appropriate evaluation methodology must be chosen, considering political, administrative, and cost constraints. In addition, we highlight the role evaluations have played in shaping employment policy in the United States. We also consider factors that influence the extent to which evaluations are performed and used and offer suggestions on how evaluations may play a larger role not only in making policy decisions but also in helping local administrators improve the performance of their operations.

Publicly financed job training in the United States

The majority of publicly financed job training programs are federal programs. Their primary purpose is to assist workers in gaining the skills necessary to find gainful employment. While it is recognized that workforce development promotes local and regional economic development, federal programs focus on the benefits to individuals and give little attention to their possible effects on local economies. Whatever benefits may accrue from these programs to local areas is only of secondary interest. State governments, on the other hand, have taken a more active role in using job training programs to promote local and regional economic development. During the past two decades, states have implemented training programs that benefit firms directly. The purpose of these programs is to enhance the business environment within their state by providing training that meets the needs of local businesses.

In this paper, we distinguish between federal programs that benefit individual workers and state programs that target firms for economic development purposes. Most rigorous evaluations of training programs have focused on the federal programs. Although these evaluations have been conducted in selected local areas, they look exclusively at the outcomes of individuals in the form of employment and compensation and do not look at the effect of these changes on the local economy. Evaluations of state programs that target economic development efforts are much less rigorous, primarily because it is more difficult politically and administratively to use random assignment design when firms and government jurisdictions are involved. Therefore, we will focus our attention on the evaluation of federal programs first and then turn to ways to evaluate the effect of training programs on local economies.

Approaches to delivering publicly financed job training

Publicly financed job training programs have pursued various strategies for delivering services. They range from providing instruction in remedial skills, such as reading and arithmetic, to offering training on detailed procedures to perform at a high level in a specific occupation. Table 11.1 provides a taxonomy of the various types of training prevalent in government programs. Job training is different from the more formal educational process in that it is usually short term and focuses on mastering specific skills. For programs targeted at individuals, the goal of job training is to address a structural mismatch between the skills of a job seeker and the needs of an employer, so that the individual can return to work. This objective may entail training on a specific type of equipment or provide basic training about proper workplace behavior or job search skills.

Training takes place in a variety of settings. The traditional approach uses the classroom setting, which can accommodate both general instruction and customized training. At the other end of the spectrum is on-the-job training. This takes place in the workplace and typically in the setting in which the worker will eventually be assigned. Of course, some training uses a combination of the two approaches, such as the case for various youth programs. Post-employment training also combines classroom activities with laboratory and related activities which are directly linked to continued employment and advancement in a specific job or occupational field.

Table 11.1. **Types of job training**

Occupational skill training
Provided in group setting is called institutional or classroom training and usually for occupations in general demand.
Customized is designed to suit the specific requests of an employer with available job slots or employees already on-board.
Vouchers are a vehicle to allow participants to choose among approved topics and training providers.
Skill training provided in an experiential workplace setting is referred to as on-the-job training (OJT).
When OJT is provided through a public agency it is sometimes called work experience.
Remedial training
General training which seeks to remedy basic gaps in reading and mathematics skills to make job seekers ready for skill training.
Classroom soft skills training
Conveys knowledge about proper workplace behavior or job search skills.
Post-employment training
Combines classroom and practical activities intended to promote retention and advancement within a given career path.
Youth training programs
Basic skills training in a workplace context, support for further general education and credentials, mentoring, school-to-work and school-to-apprenticeship transition services, intensive residential education and occupation and job training.

Recent federal job training programs in the United States

Federal job training policy has its origin in depression-era *New Deal* programs for public works in the 1930s. Renewed training efforts thirty years later were greatly influenced by new economic goals set during President Johnson's *War on Poverty*. Subsequent programs reflected political preferences toward different population groups and the economic realities of the times. A summary of the four main post-war federal job training programs is provided in Table 11.2.

Table 11.2. **A Chronology of federal job training programs in the United States**

Program	Training Types	Eligibility	Intergovernmental Relations
Manpower Development and Training Act (MDTA), 1962	Institutional and on-the-job training (OJT).	Low income and welfare recipients.	Federal funding granted directly from 12 regional offices to agencies in local areas. Administration and reporting structures similar.
Comprehensive Employment and Training Act (CETA), 1973)	On-the-job training, Classroom skill training, Classroom soft training, and Work experience in public agencies, and Public Service Employment (PSE).	Training was targeted to low income persons, welfare recipients, and disadvantaged youth.	Federal funding granted to prime sponsors in substate regions which numbered about 470. Performance monitoring with results reported to the US Department of Labor (USDOL).
Job Training Partnership Act (JTPA), 1982	On-the-job training, Classroom skill training, Classroom soft training, and Work experience in public agencies.	Low income, public assistance recipients, dislocated workers, and disadvantaged youth.	Federal funding through state governors to private industry councils (PICs) in each of 640 service delivery areas. PIC performance reports to governors who reported to USDOL.
Workforce Investment Act (WIA), 1998	On-the-job training, Customized classroom skill training, Classroom soft skills training, and Work experience in public agencies.	Access to core services like job search skills and job referral is unrestricted. Training is targeted to the most difficult to reemploy.	Like JTPA, but PICs became fewer (600) workforce investment boards (WIBs) with private sector majority membership. Monitoring is reduced relative to JTPA practice.

Source: O'Leary and Straits (2002).

Manpower development and training act

Under the Manpower Development and Training Act (MDTA) of 1962 job training was targeted to the low income and welfare recipient populations. Funds were allocated to communities based on population and estimates of the proportion below the poverty income level. The federal government

managed MDTA funding through 12 regional offices of the US Department of Labor. Localities bid for federal funding to provide training programs, which eventually led to problems of duplication of effort and to the need for coordination at higher levels of government.

Job corps

The Job Corps, established in 1964 by the Economic Opportunity Act, is a one-year residential program for disadvantaged youth. The Job Corps provides remedial academic instruction, job training, and other supportive services. It has remained largely unchanged over the years.

Comprehensive employment and training act

The 1970s brought a more comprehensive approach to addressing the problems of the economically disadvantaged. A move toward decentralization of employment programs transferred decision making authority from the federal to state and local governments. Authority as defined in the legislation and regulations often included responsibility for designing, implementing and evaluating program activities.

The Comprehensive Employment and Training Act (CETA) of 1973 introduced the concept of a local advisory board to assure that local public interest would guide program planning. The private industry council (PIC) membership and role were established in the regulations and in some localities representation was "guaranteed" for constituencies like education and labor. CETA job training was targeted to economically disadvantaged, welfare recipients and disadvantaged youth.

Job training partnership act

Under the Reagan administration in the 1980s, publicly funded job training was reoriented toward serving the needs of private-sector employers. Classroom skill-training was identified as a major weakness of prior programs, since it was often not the kind of training desired by local employers.

The Job Training Partnership Act (JTPA) of 1982 limited training choices to skills that were in demand by local employers. JTPA also increased the private sector share of members on local job training advisory committees to ensure that their interests were taken into consideration. Evaluation was an integral part of the program which was said to be performance driven through a system of performance standards for participant reemployment rates and earnings. In response to the widespread layoffs associated with economic restructuring in American business during the 1980s, JTPA job training was targeted to dislocated workers in addition to the economically disadvantaged and welfare recipients.

Workforce investment act

By the late 1990s economic conditions had improved to the point where full employment existed in most of the United States. The more than 30 years of searching for ways to reduce poverty through employment policy evolved into a new approach that shifts responsibility from government to the individual, and divests authority from the federal government to the states. It exchanges an emphasis on skill training that will lead a family out of poverty for an emphasis on job placement that will quickly reduce the cost of public assistance payment.

Reflecting these changes in policy toward self-sufficient and local control, Congress passed the Workforce Investment Act in 1998 to replace JTPA. WIA reforms federal job training programs and creates a new comprehensive workforce investment system. The reformed system is intended to be customer focused, to help individuals access the tools they need to manage their careers through information and high quality services, and to help employers find skilled workers. The new emphasis of WIA is on "work first". In other words, a job is the best training. If jobs are not available, training will mostly be customized to serve employer needs, on-the-job training, and short term training in core skills.

Key innovations brought by WIA are: 1) one-stop career centers where all employment and training programs are assembled in one physical location, 2) individual training accounts which act as vouchers for job seekers requiring skills improvement for labor market success, 3) universal access to core employment services with sequential, more restricted access to intensive services and training, and 4) accountability monitored through performance indicators. A significant feature of WIA for local areas is the increased private sector control over use of training funds. Workforce Development Boards (WDBs) are to have a significant majority membership from the employer community. Targeting the most difficult to reemploy and follow-up monitoring of outcomes were retained from JTPA.

Who gets job training?

According to OECD comparative statistics, the United States spends about 0.04 per cent of GDP on job training programs. As shown in Table 11.3, this proportion is low compared to other industrialized nations, placing the United States in the lowest fifth of the countries included in the list. Job training comprises 26.7 per cent of US expenditures on all active labor market programs (ALMPs). This percentage is comparable to that of Germany and higher than that of France and the United Kingdom.[1]

Table 11.4 shows both public and private expenditures on job training. Government financed job training comprises about 11 per cent of the $68 billion

Table 11.3. **Government expenditures on job training as a percentage of GDP in 25 OECD Countries, 2000**

	As a percentage of GDP			Training as a percentage of spending on	
	Training	ALMPs	LMPs	ALMPs	LMPs
Denmark	0.84	1.55	4.51	54.2	18.6
Finland	0.35	1.07	3.29	32.7	10.6
Germany	0.34	1.23	3.12	27.6	10.9
Sweden	0.31	1.38	2.72	22.5	11.4
Netherlands	0.30	1.57	3.65	19.1	8.2
Portugal**	0.30	0.51	1.34	58.8	22.4
Spain	0.29	0.84	2.18	34.5	13.3
France	0.28	1.36	3.12	20.6	9.0
Belgium*	0.25	1.36	3.70	18.4	6.8
New Zealand	0.18	0.55	2.17	32.7	8.3
Austria	0.17	0.49	1.58	34.7	10.8
Canada	0.17	0.51	1.49	33.3	11.4
Greece**	0.17	0.35	0.83	48.6	20.5
Italy*	0.12	0.63	1.28	19.1	9.4
Korea	0.09	0.46	0.55	19.6	16.4
Switzerland	0.09	0.48	1.05	18.8	8.6
Norway	0.08	0.77	1.16	10.4	6.9
Hungary	0.07	0.40	0.88	17.5	8.0
United Kingdom	0.05	0.36	0.94	13.9	5.3
Mexico*	0.04	0.08	0.08	50.0	50.0
United States	0.04	0.15	0.38	26.7	10.5
Japan	0.03	0.28	0.82	10.7	3.7
Australia	0.02	0.45	1.50	4.4	1.3
Czech Republic	0.02	0.22	0.52	9.1	3.9
Poland	0.01	0.15	0.96	6.7	1.0

* 1999.
** 1998.
Where GDP is gross domestic product, ALMP is active labor market policies, and LMP is labor market policies.
No data available for OECD countries: Iceland, Ireland, Luxembourg, Slovak Republic, and Turkey.
Source: OECD (2001).

that was spent for that purpose in fiscal year 2001. The federal government accounted for nearly 90 per cent of the public expenditures. Of the $6.4 billion in federal expenditures on job training, 39.6 per cent went to adult disadvantaged and dislocated workers, 43.3 per cent to youth programs (Job Corps and others), 6.9 per cent to community service employment for older workers, and 2.1 per cent to workers impacted by changing patterns of international trade.

Background characteristics for participants in the three main federally funded employment and training programs are summarized in Table 11.5. Differences in the characteristics of participants within the various programs

Table 11.4. **Estimated expenditures for public job training programs in the US, fiscal year 2001**

Thousands of dollars

Programs	Federal funding	Share of federal funding %	State supplemental funding	State financed customized FY 1998	Employer financed 1998	Grand total funding
Adult and dislocated worker activities	$2 540 040	39.6				
Youth activities	$1 377 965	21.5				
Job corps (youth)	$1 399 148	21.8				
National Programs	$528 150	8.2				
Other programs (non-WIA)	$4 500	0.1				
TAA training	$94 400	1.5				
NAFTA training	$37 150	0.6				
Community service employment for older Americans	$440 200	6.9				
Total funding	$6 421 553	100.0	$276 621	$593 191	$60 700 000	$67 991 365
Percentage of grand total of funding	9.4%		0.4%	0.9£	89.3%	100.0%

WIA – Workforce Investment Act.
TAA – Trade Adjustment Assistance.
NAFTA – North American Free Trade Act.
Source: Wandner, Balducchi, and Spickard (2001).

are consistent with the groups these programs are intended to serve. Both JTPA Title II-A and Title III programs were intended for adults, but Title II-A was for the disadvantaged defined as having income below the poverty line, whereas Title III was for dislocated workers who were unable to find work and needed additional training. Consequently, as shown in the table, a larger percentage of participants in Title II-A was welfare recipients and fewer were high school graduates than those in Title III training programs. More women participated in Title II-A than in Title III. The youth training program (JTPA Title II-C) obviously had a preponderance of participants who had not finished high school.

The bottom of Table 11.5 provides gross outcome information for participants in the three major JTPA funded programs. Entered employment was 68 and 69 per cent for the adult and dislocated worker programs, respectively, while it was 47 per cent for the youth program. For youth, sizeable proportions also achieved an employment enhancement or competency, which JTPA also regards as success. Among those entering employment at program exit, hourly earnings rates were estimated to be $8.75, $7.07, and $11.95 for adult, youth, and dislocated workers, respectively.

Table 11.5. **Characteristics and outcomes of JTPA training participants, PY 1999**

Characteristics	Adult Title II-A	Youth Title II-C	Dislocated workers Title III
Number of program participants	133 774	58 548	189 794
Gender: Female (%)	65	58	54
Age: 14 to 15 (%)		7	
Age: 16 to 21 (%)		93	
Age: 22 to 54 (%)	97		89
Age: over 55 (%)	3		11
Education: less than high school (%)	22	71	11
Education: high school (%)	56	26	50
Education: post high school (%)	22	3	39
Race: black (%)	35	34	19
Race: Hispanic origin (%)	16	23	13
Race: while (%)	43	38	62
Disabled individual (%)	7	12	2
Welfare recipient (%)	26	19	2
Ex-offender (%)	18	13	5
UI recipient (%)	10	1	69
UI exhaustee (%)	3	1	5
Veteran (%)	6		11
Outcomes:			
Entered employment rate (%)	68	47	69
Average hourly wage ($)	8.75	7.07	11.95

Evaluation of job training in the United States

The two most popular assessment techniques for job training programs are *performance monitoring,* which tracks gross outcomes, and *net impact estimation,* which assesses the incremental value of an intervention. Each of these approaches has advantages and shortcomings (Barnow and Smith [2002], King [2002]). The focus of this paper is on net impact estimation of job training impacts. Such estimates of the incremental value of a program treatment are the basis for net benefit computations to estimate the return on investment for government expenditure.

Net impacts compare mean outcomes of a representative sample of program participants to an appropriately chosen sample of similar persons who did not receive services. Great care must be taken in forming the latter group which is called the counterfactual.[2] The difference between the two groups on outcome measures of interest is the estimate of the program effect. For the JTPA program, Congress stated that success would "be measured by the increased employment and earnings of participants and the reductions in welfare dependency" (Barnow [1989], p. 117).

Methods for estimating job training impacts

Evaluation approach

Net impact estimation can be based on samples gathered through classically designed experiments involving random assignment, or through quasi-experimental studies that use statistical means to mimic the ideal of an experiment. In a classically designed experiment, the participant and comparison groups are created by random trials. This means that fixed experimental conditions are repeated a sufficient number of times to generate the required sample sizes, and on each repetition a random assignment is made either to program participation or to a comparison group.[3]

Quasi-experiments attempt to mimic a classical experiment by creating appropriate participant and comparison groups using statistical means instead of through random assignment. An expedient approach is to use the prior experience of job training participants as a comparison group for themselves. The associated net impact estimator is called the pre- *versus* post-program participation estimator. This approach is inexpensive because it either relies on administrative data or cuts follow-up interview numbers in half. It implies that using participants as their own comparison group automatically adjusts impact estimates for both observable and unobservable differences in characteristics. This strategy is acceptable if the evaluation budget will not support another approach, but there are intrinsic problems.

Ashenfelter (1978), in evaluating retraining programs under MDTA in the United States, found evidence of an earnings dip prior to job separation for dislocated workers. Evaluating the net program impact on re-employment earnings using the pre-post design will lead to an over-estimate of net program impacts due to the presence of the "Ashenfelter dip".

An alternative to pre-post analysis is to select a contemporaneous comparison group by matching on observable characteristics. This approach selects a comparison group of observationally similar people who became jobless about the same time as the job training participants but did not enter the program. For each job training participant, a "twin" is selected by matching on variables such as age, gender, race, educational attainment, prior occupation, prior industry, average prior earnings level, prior receipt of unemployment compensation, prior use of active labor programs, and the geographic location of residence.[4]

If some characteristics differentiating participants and non-participants are not observable, it may be possible to indirectly account for them through propensity score matching. A model predicting whether or not someone participates in job training is estimated on observable characteristics, some of which explain participation but do not explain re-employment success. Such

factors are difficult to come by so such models of participation are most often identified by using non-linear transformations of observable variables. Once the participation model is estimated on the whole sample, it can be used to create a comparison group by selecting non-participants who have a participation score closest to each participant.

Estimation techniques

If random assignment has been properly implemented in the context of a field experiment, then the net impact can be estimated as a simple difference in means of the treatment and comparison groups. After validating homogeneity of treatment and control groups, the randomly assigned samples can be used to compute net impact estimates by a simple difference of means:

[1] $E(y_p) - E(y_c)$,

where E is the expectation operator yielding means of the random variables, y is an outcome of interest, and the index p denotes the sample of job training participants and c denotes the comparison sample. T-statistics are used to test for statistical significance.

In terms of clearly guiding policy, simple unadjusted net impact estimates based on random trials are usually the most influential because they are easy to understand.[5] An equivalent approach is to use regression analysis to compare the outcomes of the treatment and comparison groups. Program impacts can be estimated by running the ordinary least squares model:

[2] $y_i = a_0 + a_1P_i + u_i$,

on a pooled sample of comparison group members and job training participants, where y is the outcome of interest, a_1 is the impact of the program on the outcome for the job training participants, a_0 is the mean value of the outcome for comparison group members, P is a dummy variable with a value of 1 for job training participants and 0 otherwise, u_i is a normally distributed mean zero error term, and i is an index denoting individuals in either the participant or comparison group samples. Tests for significance of program impacts are simply t-tests on the parameter a_1.

For most of the quasi-experimental approaches *regression adjustment based on observables* is the expedient and satisfactory net impact estimation technique. This can be applied whether comparison groups are created using pre-versus post-program participation samples, or gathered as contemporaneous non-participants. This is also a useful approach for using data generated by a field experiment when complete homogeneity of groups does not result. Computationally the method involves a simple extension of equation [2]. In such cases, estimation of the model:

[3] $y_i = a_0 + a_1P_i + b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni} + u_i$,

by ordinary least squares on the pooled sample of participant and comparison group members yields net program impact estimates.[6] Equation [3] is the same as equation [2] except that the observable characteristics (X_i) of participants and non-participants are included.[7]

This method yield net program impacts adjusted for observable characteristics. The estimates are called "net" because the comparison and program participant groups are statistically adjusted so as to remove heterogeneity across the samples. That is, the only remaining factor contributing to a difference in the outcome measure is exposure to the program treatment. The estimation methodology nets out all other observable factors affecting the outcome.

When the comparison group has been created using a matched-pairs algorithm, net program impacts may be estimated simply as a difference in means on matched pairs.[8] That is applying the technique of equation [1] to the data. Some analysts have also applied the method of equation [3] to such data where the co-variates (X) include both the matching characteristics and some additional variables.

The popular approach of difference-in-differences involves preparing pre-post data on both job training participants and a contemporaneous group of program non-participants.[9] For all subjects in each group the difference between the value of the most recent outcome and the value of an outcome prior to participation in the program is computed. This variable is the first difference. The next step is to apply the technique of equation [1] to the data on this variable. Thereby the second difference is computed as the contrast in the first differences between job training participants and non-participants. The motivation for this approach is the recognition that the prior value of the outcome is an exogenous variable which embodies all the observable and unobservable ways in which the two groups differ. In principle the difference-in-difference approach automatically adjusts for the "Ashenfelter dip" in earnings.

When selection into programs is not random and participation in a program is due to both observable and unobservable characteristics, program impacts cannot be properly estimated in a regression model of the type specified in equation [3]. Heckman (1979) showed that sample selection will bias parameter estimates computed by OLS in an equation like [3] as if an important variable had been omitted from the specification of the estimating equation. He recommended a way to create this omitted variable by adjusting the regression equation based on observable and unobservable factors.[10] The reason for estimating impacts using the Heckman sample selection procedure is the concern that there is something unobservable about program non-participants who have observable characteristics similar to job training participants, which

would cause them to have different labor market success than job training participants even if they had participated in the same program.

Criteria for good estimation methods

Table 11.6 lists seven criteria for assessing whether or not an estimation method is a good one for the task at hand. The seven criteria were listed in the table in their approximate order of importance.

The first criterion is *transparency* of impact estimation methodology. By this we mean that the important users of the evaluation results can clearly understand the method of estimation. Program impact estimates will certainly be used by senior policy makers who are relatively knowledgeable about the methods of social science research. Other users will be less comfortable with complex statistical procedures. The estimates should be of interest to those bodies which decide the allocation of national employment funds. Ideally the estimates will also be used by program managers at the regional level, case workers at the local level, and even potential job training participants. The estimates will carry more weight if it is clear how they were computed.

The classically designed field experiment satisfies the criteria of transparency better than other methods. A simple comparison of means between treatment and control groups can be very persuasive, if the two groups are shown to be randomly assigned and homogeneous. One must recognize, however, that experimental evaluations may be prohibitively expensive. A

Table 11.6. **Design elements in a net impact evaluation**

Evaluation approach:
Classically designed experiment
• Random trials
Quasi-experiment
• Pre-versus post-program participation
• Matching on observable characteristics
• Matching on observable and unobservable characteristics
Estimation techniques:
Difference in means
Regression adjustment based on observable variables
Difference in means on matched pairs samples
Difference in differences
Regression adjustment based on observable and unobservable variables
Criteria for good estimation methods:
Transparency of impact estimation method
Unbiased
Efficient
Robustness
Insensitivity to small sample sizes
Insensitivity to small variations in estimator
Reproducibility

matched-pairs approach to comparison groups can be nearly as convincing as an experiment, if explained clearly. Least desirable is a complex statistical method which may be required in situations where samples of participant and comparison group people drawn are widely dissimilar.

Given an appropriate sample, an *unbiased* estimator will yield an impact estimate reflective of the benefits to the population of job training participants. If the estimator is *efficient,* then there will be little variation in values of the estimated program impact across alternative samples. If the estimator is *robust,* it is not greatly affected if some observations have values far from the norm.[11] Having an estimator which is relatively *insensitive to small sample sizes* permits broad latitude for sub-group analyses on demographic and program design features.

Insensitivity to small variations in the estimator concerns the analyst's ability to control for observable and unobservable factors when computing net impacts. In practice it is impossible to control for all factors. With an estimator which is insensitive to small variations, by controlling for the most important factors we should be insulated from other influences.

Reproducibility is a desirable feature in any scientific investigation. The data sources should be well documented and the statistical methodology well known so that another investigator looking at similar evidence would arrive at a similar conclusion about the population parameters of interest.

The two considerations which are noticeably absent from our list of criteria for good estimators are timeliness and cost effectiveness. These are both practical considerations for implementation of an evaluation approach. Users of information from program evaluations have decision time lines which can influence the selection of an approach, and they have more or less limited budgets for doing evaluation which also can influence the decision. The main distinction in choosing the evaluation approach is the experimental *versus* quasi-experimental choice.

The simple difference of means estimator which is possible given the experimental approach to evaluation satisfies all our criteria for a good estimator. Unfortunately, the time and budget are not always sufficient to support the experimental approach. Most often data for the evaluation is gathered by a quasi-experimental design and participant and comparison samples are selected either with a pre-post design or a contemporaneous approach. With the contemporaneous approach there can be extensive matching *ex ante* on observable characteristics or there can be simple eligibility screens imposed. For quasi-experimental designs the last approach is most general. Contemporaneous samples with inflow eligibility screens applied will permit application of most estimation techniques and will therefore permit selection of the best by applying the criteria we have discussed.

Practical issues in evaluating job training

The most practical aspect of any evaluation design is the plan for data collection. This component, which may be called the sample design, has five essential elements: 1) setting the sample size, 2) determining the geographic sites covered by the evaluation, 3) establishing procedures for drawing the sample, 4) survey implementation, and 5) preliminary examination of the data.

Sample size

The samples of participants and non-participants should be of sufficient size to ensure precision of desired impact estimates. Larger sample sizes permit detection of smaller program effects while simultaneously increasing the precision of estimates as measured by sampling errors.[12] The sample sizes for net impact evaluations should be big enough to reliably measure effects of a size that would be of interest to policy makers. For example, if program were unacceptable due to generating anything less than a 15 percentage point increase in the probability of re-employment, then the sample size should be set to detect an impact at least that small. The constraining factor in collecting sufficiently large samples to ensure precise estimates for any expected effect is the cost of collecting data.

Table 11.7. **Practical issues in evaluating job training**

Preliminaries:
Sample design
Randomly select samples of persons for participant and comparison groups
Extract records from existing administrative records on samples selected
Prepare a data file for preliminary analysis of samples selected
Prepare lists of names for interviews
Survey work:
Pilot test surveys
Revise surveys and set final formats for recording survey responses
Prepare surveys in format required for interviews
Prepare a training manual for survey workers to conduct interviews
Designate survey managers for major geographic regions
Assemble a team of survey workers to conduct interviews
Conduct survey worker training
Conduct interviews with established call back protocol
Deliver completed questionnaires for data entry
Final Data Processing:
Error checking, correction, and key entry of data to computer files
Preparation of computer files for data analysis
Delivery of data files to data analysts
Correction of data files based on questions from data analysts
Impact Estimation

Source: Adapted from Table 4.13 in O'Leary, Nesporova and Samorodov (2001, p. 106).

Site selection

All training program evaluations are conducted at local sites, since training for the most part is site specific. In conducting a net impact evaluation involving follow-up surveys, practical consideration must be given to the mode of survey. Telephone surveys involve fewer geographic considerations than house-to-house surveys. For in-person surveys, transportation costs are a major component of the survey costs, particularly in areas with significant rural populations. A standard requirement in US evaluations is that sites be selected to be representative of the state or region where the evaluation is being conducted.

Sample selection

In classically designed field experiments, randomization takes place at the time of program assignment within a local employment office. Ideally, sample selection operates as an inflow process for quasi-experimental evaluations too. Both job training participants and comparison group members should be in the same labor market status and should both have the same eligibility for program participation.

For quasi-experimental evaluations, drawing program participant and comparison group samples from the data systems of government employment offices permits economizing on the length of any necessary follow-up surveys. In most countries it will be possible to draw important baseline information from the computerized data systems of employment offices which are used for program administration. Such systems can often provide reliable basic demographic information about things like age, gender, educational attainment, and family composition. Other information like usual occupation and industry of employment, prior average monthly earnings, and prior use of government employment and income support programs is also potentially available. Gathering such data from administrative records improves accuracy and shortens follow-up surveys to focus on re-employment experiences.

Survey implementation

Survey researchers have found that a response rate of about 80 per cent is required to ensure that the data gathered reflects the population of the sampling frame. An expedient approach to conducting the field work of surveys is to contract with a private survey company. Using a private independent company that is a disinterested observer also adds a level of objectivity to the process which may help to validate the findings. However, the cost per survey completed using such an outside contractor may be prohibitive for some programs, particularly if financed by local governments. An alternative approach is to develop survey research capacity within the system of local job training administrators. Some local administrative entities already have the

responsibility of following up with past participants to measure outcomes, and this capability could be extended to evaluations.

The plan should require a complete canvas of all listed names on the first attempt to contact. The number of interviews completed in the first round should then be compared to the sample target. If the numbers fall significantly short of the target then a second call attempt should be made for all subjects not contacted in the first round. This process should be completed until the sample target is made. By this approach the final sample will be either slightly under or somewhat over the design number. This strategy preserves the representativeness of the sample such that each person in the original sampling frame has an equal chance of being in the final sample.

One canvassing plan which absolutely should not be followed is to stop all interviews immediately after completing the targeted number of interviews. A round of interviews should always be completed each time one is started. Interrupting a round will diminish the representativeness of the sample.

In comparing the cost of the alternative approaches to survey field work all aspects of the project should be considered. Once the sample has been drawn, many steps are required before a final data file is available for analysis. In their usual chronological sequence, the main steps are: 1) training survey workers, 2) pilot testing the questionnaires, 3) revising questionnaires, 4) printing questionnaires, 5) distributing interview address lists and questionnaire copies to survey workers, 6) maintaining records of multiple call back attempts, 7) supervising accuracy and completeness, 8) computer key entry of survey data gathered, and 9) error checking the computer files of survey data.

Preliminary examination of the data

With a quasi-experimental design, data for evaluation will usually come from two sources: administrative records and follow-up surveys. The first step in evaluation is to examine the data for completeness and errors. Summary information for all variables should be printed out and examined. The summary information should include the mean, minimum, maximum and a count of the number of observations with missing data. All of these should be checked to see that they fall in the range of acceptable values. Accepted procedures are then followed to correct for unreasonable outliers and missing values.[13]

In addition, it is necessary to check for randomization in the creation of program participant and comparison groups.

Examples of federal job training evaluations

Each of the major federal job training programs has been evaluated in one form or other. The evaluation of MDTA examined gross outcomes based upon a survey of previous participants. Improvements were made when CETA was

up for evaluation by including a comparison group. The evaluation of JTPA took the next step to adopt a random assignment design. Job Corps was subject to both a quasi-experimental and random assignment evaluation. An evaluation of WIA based a quasi-experimental approach using administrative data is currently underway. Since JTPA and Job Corps are the only two programs that used random assignment evaluation design, most of what we know about the effectiveness of job training programs is gleaned from evaluations of these two programs. We also note in our description of these evaluations the role evaluations played in the political process of legislative reauthorization.

Manpower development training act

The MDTA was the first federal attempt to help displaced workers find reemployment through job skill retraining. It addressed the main concern at the time of job loss due to technological change (Leigh, 1990). Between September 1962 and September 1967, more than 601 000 people were enrolled in retraining programs organized by local areas that received federal grants directly from regional offices of the US Department of Labor.

The evaluation of MDTA was based on follow-up surveys of participants in the program. At the time of the major MDTA evaluation, 74 000 participants were still involved in retraining programs and 30 per cent had dropped out. Among earlier participants, 90 per cent had obtained reemployment during the year after training, and 77 per cent were employed at the time of the last follow-up survey (Mangum 1968, p. 81). As mentioned in the previous section, however, these gross outcome estimates are not reliable indicators of net program impacts.

Sunset provisions in the MDTA legislation ended the program in 1969 and obviously evidence from the evaluation did not prevent the program from being terminated. The prime reasons for the demise of MDTA were the administrative structure whereby the authority of state and local political entities was circumvented with federal contracts going directly to local service providers, and the delivery of services was duplicated at the local level.

CETA

The Comprehensive Employment and Training Act (CETA) of 1973 was the first training program for which the US Department of Labor developed a data base specifically intended for program evaluation (Leigh 1990, p. 10). It was called the Continuous Longitudinal Manpower Survey (CLMS) and contained data on program participants, on comparison group members drawn from the national labor force survey (Current Population Survey), and on earnings for all subjects from national social insurance (Social Security) records. Evaluation

studies were greatly facilitated by the creation of CLMS, despite the fact that CETA programs were targeted to low-income individuals while the labor force survey represented the nation.

Three main findings emerged from 11 major CETA evaluations (Leigh 1990, p. 11). First, there were no measurable employment or earnings impacts for men. Impacts for women, on the other hand, were positive and significant. Second, on-the-job training was usually more effective than classroom training. Finally, the range of impact estimates was quite wide, despite the fact that all analysts used the same CLMS data.

CETA expired in 1982. However, its termination was the result of issues surrounding the administration of the program and mismanagement of funds more than the evidence of its effect on workers.

JTPA

The passage of JTPA was the result of true ideological and partisan compromise between liberal Democrats and conservative Republicans. Many features of the bill reflected the compromise needed for its enactment, including a Congressional mandate for a major national evaluation of the program. To assure an objective net impact evaluation, Congress insisted that the evaluation be based on methods of field experimentation with random assignment of subjects to training and to comparison groups in 16 sites across the country. Orr *et al.* (1996, p. 109) report that training to economically disadvantaged adults resulted in 11 per cent greater earnings for women and 6.7 per cent greater earnings for men. For both genders, the earnings gains were mainly due to increases in hours worked. Both men and women experienced positive net benefits, and the net benefit to society for both genders was just over $500 per participant (Orr *et al.* [1996], p. 189).

Job Corps

The first major evaluation of Job Corps was quasi-experimental (Mallar *et al.* [1980]) and found modest positive effects on employment and weekly earnings, but no impact on hourly wage rates. A recent study was done as a classically designed field experiment (Burghardt *et al.* [2001]). The new study found that Job Corps participation results in significant earnings gains for disadvantaged youth. "Furthermore, earnings gains, educational progress, and other positive changes were found across most groups of participants and are expected to persist as they get older."[14] It is reported that the most recent study was instrumental in saving Job Corps from being eliminated in the latest budget rounds.

WIA

Although the WIA was enacted in 1998, full implementation of the WIA did not begin until July of 2000. Consequently, only recently has enough time passed for the full impact of the program to take place. Work has just begun on a seven state quasi-experimental net impact study of training (Hollenbeck 2002). The states involved are Washington, Missouri, Illinois, Maryland, Florida, Georgia and Texas. A field experiment to evaluate training vouchers under WIA is also in progress (Decker and Perez-Johnson [2002]).

As a precursor to this larger evaluation, Hollenbeck used the same method to evaluate training programs in the state of Washington. Hollenbeck (2002) used this non-experimental approach of statistical matching to evaluate workforce development programs in the state of Washington.[15]Using administrative data collected by the various state agencies, Hollenbeck used applicants to the state's employment service as the source of individuals to whom he would compare the participants in the training programs. Next, he used statistical matching on an individual-by-individual basis to find the people who most closely matched program participants in terms of observable characteristics. Net impacts were then determined by comparing outcomes for individuals who participation in the training programs to their matched counterparts from the employment service data, who never participated in any programs. Using this method, Hollenbeck evaluated nine job training programs in the state, including those associated with community colleges. One benefit to this approach is that the data are already collected by the various agencies and readily available for such an evaluation, eliminating the need to collect additional data which is typically the case for random assignment experiments.

What works for whom?

Evaluations of job training in the United States have been conducted by type of service and category of participant targeted by federally funded programs. The main target groups include disadvantaged adults and youth and dislocated workers. Within these target groups, different results have emerged by gender and whether job training is classroom or experiential.[16]

Representative of the common pattern of job training impact estimates are results from two national evaluations conducted as field experiments involving random assignment. We illustrate the pattern of findings with evidence from these two evaluations: the national JTPA evaluation (Orr *et al.* [1996]) and the Job Corps evaluation (Burghardt *et al.* [2001]). It is noteworthy that evidence from the JTPA and Job Corps experiments is consistent with other results, since these two evaluations both applied a well defined counterfactual. When the treatment is job training, most evaluations require

that the control group be excluded from all services. The JTPA and Job Corps evaluations define a counterfactual permitting that the control group may have received any number of services other than the publicly financed job training being evaluated. This design provides better external validity, meaning the experimental evidence applies more naturally to the real world context.

Among disadvantaged adults, the measurable impacts for men tend to be small. Impacts for women, on the other hand, tend to be positive and significantly larger than for men. Among women welfare recipients, earnings impacts are also significant and positive. For youth, short-term job training tends to be ineffective, whereas the longer duration, more intensive Job Corps type training appears to be effective for disadvantaged youth. Furthermore, on-the-job training (OJT) is usually more effective than classroom training (CT).

Evidence from the national JTPA evaluation is summarized in Table 11.8. It shows earnings impacts over the 30-month period following program entry. Adult male JTPA training participants on average gained $1 599 in earnings and tended to benefit more from on-the-job training (OJT) than from classroom training (CT).

Table 11.8. **Impacts of job training by program and demographic group**

Demographic group program	Mean Impact on earnings	Standard error
Adult men		
JTPA	$1 599*	86 565
OJT	$2 109	133 535
CT	$1 287	158 282
Adult women		
JTPA	$1 837**	52 525
OJT	$2 292**	102 323
CT	$630	67 070
Adult welfare women		
JTPA OJT	$4 833**	172 929
JTPA CT	107 777	106 262
Youth Male		
JTPA OJT	-$3 012	222 222
JTPA CT	$251	191 616
Youth female		
JTPA OJT	-$579	188 383
JTPA CT	$839	79 191

* Statistically significant at the 90 per cent level in a two-tailed test.
** Statistically significant at the 95 per cent level in a two-tailed test.

Source: Orr *et al.* (1996). Notes: Earnings impacts are estimated over 30 months after program entry, JTPA is Job Training Partnership Act, OJT is on-the-job training under JTPA, CT is classroom training under JTPA.

JTPA impacts for adult women tend to be positive and statistically significant. The point impact is $1 837 over 30 months, with OJT significantly more effective than CT. For women welfare recipients in JTPA, OJT had a positive impact of $4 833 on earnings, and this was significantly greater than the impact for CT for this group.

For youth, the short-term job training offered by JTPA was not efficacious; however the expensive and long-duration Job Corps training had a sizeable positive impact. Job Corps had per-participant net benefits to society of nearly $17 000 over the lifetime of program participants (Burghardt *et al*. [2001]). This result for Job Corps youth is the combination of several effects: additional hours of work, higher rate of pay, and less social assistance.

Overall impact of federal job training programs

As mentioned in a previous section, expenditures on publicly provided job training programs are a small percentage of the estimated total amount that private-sector employers spend to train their workforce. Only 10 per cent of the estimated $68 billion dollars spent on training in 2001 was financed by either the federal or state governments. The rest was financed by businesses. Moreover, these programs served only a fraction of the eligible workers. For instance, a US Government Accounting Office Report estimated that during the late 1980s JTPA served no more than 6 per cent of the total estimated eligible population.[17] This percentage dropped even lower near the end of the program in the late 1990s.

Therefore, the question arises as to the net impact of the program on the targeted population beyond its effect on those who actually received services. The issue is a matter of scale and the stated objective of the program. If full employment is the goal (as measured by the unemployment rate for a region or nation), then the question is how much the program should be expanded in terms of both coverage and intensity of services in order to accomplish that objective. The goal of full employment of a region or targeted group could be achieved, undoubtedly with considerable more financial commitment, without making much progress in moving people above the poverty line. If reducing poverty is also the goal, then it is conceivable that the program must be expanded even further. For example, Heckman, Roselius and Smith (1994) estimate that in order to restore high school graduates in 1989 to their 1979 wage levels, it would require $212 billion (in 1989 dollars) and another $214 billion to restore male high school dropouts to their previous wage levels, assuming a 10 per cent return from training. Even the most generous accounting of annual government expenditures on worker training pales in comparison to that amount.

State-financed job training programs targeted at firms

A large number of states subsidize the training of incumbent workers by providing customized training programs for businesses. These state programs differ from the federal programs discussed in the previous sections in that their primary purpose is to promote local and regional economy development by promoting the attraction and retention of businesses. Most states finance the training through their general fund, but some depend upon other sources such as UI-associated taxes or bonds. In 1999, the 45 states that offer such programs spent nearly $600 million on customized training programs, which accounts for about 20 per cent of the total WIA expenditures on training for adults and dislocated workers (Moore *et al*. forthcoming). From an economic development perspective, these state training programs comprised 34 per cent of the $5.2 billion spent in 2000 by states and local governments for economic development activities.

State training programs differ by size and type. Expenditures on customized training range from $117 million in California to half a million in Vermont (Moore *et al*. forthcoming). While most programs focus on incumbent workers, a few target the disadvantaged or dislocated workers. The State of Texas, for example, administers both types of programs. The Skill Development Fund assists businesses with their workforce training needs. In partnership with public community and technical colleges and a higher education extension agency, the Skills Development Fund assists in financing customized job training programs to fit the expressed needs of businesses within the state. The Self-Sufficiency Fund targets recipients of TANF and food stamps in order to assist them in obtaining training, finding a job, and becoming independent of government financial assistance. Thus, the second program, while still trying to meet the workforce needs of business, is focused more on a targeted population group. Each program spent about $12 million dollars in 2001, and combined they served more than 450 businesses that offered training to 15 000 workers. Training under the first program cost slightly less than $1 000 per participant, while training under the second program cost more than $3 000 per trainee.[18]

Evaluation methodology

Since the purpose of state customized training programs is to promote local economic development, evaluating the impact these programs requires more than simply adding up the individual effects of participants in a specific program. One must also take into account the complex relationships and interactions among local economic entities in order to assess the impact of training on the local economy. In particular, one must include in the evaluation the effect of training on productivity and wages and take into

account scale and displacement effects of the programs. Bartik, in his paper for this conference, examines various approaches and concludes that "rigorous evaluations can be done through random experimentation, statistical analysis of program users and comparison groups, surveys and focus groups, and linking regional econometric models with fiscal impact and local labor market models".

Nonetheless, it needs to be recognized that evaluations of the regional effect of state job training programs are more difficult to conduct and tend to be less precise than evaluations of training programs that are focused on individual workers. Such evaluations are asked to consider broad regional indicators, such as job creation or poverty reduction, which are the result of relatively small interventions. Rigorous evaluations must also incorporate scale and displacement effects, which are difficult to model and estimate. Random assignment methodology is possible, but it is difficult to apply to the evaluation of state programs because state job training programs are targeted to specific firms. The best one can do in many cases is to compare the desired economic outcomes of firms participating in the program with those not participating. But as discussed above, this introduces additional imprecision into the evaluation. Therefore, those contemplating an evaluation of state job training programs, as well as other locally administered social programs, approach the possibility with diminished expectations about the usefulness of the effort.

Three basic types of evaluation methodologies will be considered in this section. The first is based on a comparison group. The second uses econometric models to explicitly model the key relationships between the intervention and the intended outcomes. The third approach, a variant of the second methodology, uses a representative firm approach to trace the relative effects of programs on a firm's financial status.

Constructing comparison groups

Since state training programs typically target firms, the selection of comparison groups must be based on firms not individuals. An example of this approach is the study by Holzer *et al.* (1993), which used a quasi-experimental approach to evaluate a program in Michigan that provided one-time training grants to eligible firms. They constructed a comparison group by capitalizing on the fact that not all firms that applied for the program were successful in receiving funds. They compared the effects of unsuccessful firms to those of the successful firms (treatment group), assuming that both had the motivation to participate in the program. Their results showed that those firms that received the training grants significantly increased the amount of training they offered to their employees. The study also found that those firms that received the training grants experienced an improvement in productivity.

They also found in a separate analysis little effect of the training program on wages and only a modest increase in employment by firms.

Hollenbeck (2002) uses administrative data from employment programs in the state of Washington to conduct a quasi-experimental evaluation. The comparison group is formed by statistically matching participants with non-participants according to observable characteristics. All the job training programs and educational programs he evaluated yielded a positive net employment impact, except for community college adult basic education. While the programs that Hollenbeck evaluated did not include customized training, his study is instructive in showing how administrative data can be used to form the comparison groups.

Accounting for factors that affect economic development

Another approach to evaluating the effect of state training programs on economic development outcomes is to examine the various components that link training programs to changes in the cost advantage of firms and thus their effect on the local economy. These factors may include:

1. the effects of the program on the wages and productivity of individual workers;
2. the effect of programs on the resulting productivity and unit labor costs of firms, and
3. subsequently their effect on the location behavior of firms and the growth.

Training may have indirect effects as well. For example, businesses may view communities that invest in workforce quality as a desirable place to locate, since it signals that the community values other less tangible aspects that contribute to a favorable business climate. Furthermore, these effects must be understood within the context of local labor market conditions and with respect to the scale of the program, that is, the number of participants in these programs relative to the number of employees in the firm and the number of potential participants in the community. One must also consider the displacement effect of higher wages on the workers and on firms, which could net out any positive effects of the programs.

Once estimates of the training impacts related to these individual factors are obtained, they can be placed in a regional econometric model. One such model is REMI (Regional Economic Models, Inc.), which has been presented at this conference. The model allows one to estimate the direct and indirect effects of policy interventions such as training programs.

A simpler alternative is to consider the effects of wage and productivity effects on unit labor costs and then determine whether these programs affect the operating costs of firms sufficiently to influence their location decisions.[19]

An advantage of this approach is its simplicity and transparency. The drawback is also its simplicity, since there may be important interactions between the program and other aspects of the local economy that go unnoticed unless captured in a rigorous econometric model.

An evaluation combining comparison groups and regional econometric models

An evaluation of California's job training program, by Moore, Blake, Phillips, and McConaughy (forthcoming), combines the comparison group method with the regional modeling approach. The evaluators constructed comparison groups comprised of firms similar to those receiving training grants and then linked the difference in outcomes between the two groups to the economy through multiplier and displacement effects. The evaluators augmented the net impact analysis with case studies in order to learn more about why the outcomes differed between the two groups.

California's Employment Training Panel (ETP) was designed to increase productivity of existing firms in the state, attract new businesses, and thus reduce unemployment. The program, in essence, established a partnership between the state and participating companies to train their workers. The state reimbursed employers for the direct cost of training new hires and incumbents. In return, the company paid for the training facilities and covered the workers' salaries and the cost of lost production while workers attended training. In 1994-95, California spent $73 million on ETP with the average cost per trainee ranging from $150 to $260 per hour of class time. The evaluators surveyed firms that received the subsidies and used Unemployment Insurance administrative records to form the comparison groups.

During the period of the evaluation from 1994-95, more than 57 000 workers were trained. The most common type of ETP training contract was one in which ETP contracted with a single company to provide training for its workers. The second most prevalent approach was for ETP to contract with training agencies, such as community colleges, private training organisations, or trade associations, to provide specific types of training to workers in ETP-eligible firms. Under this contractual arrangement, the training agencies bore the risk that recruited trainees might not complete the required placement and retention on the job of 90 days, in which case the trainer would not be paid. The evaluators found that the typical firm receiving ETP training for its workers employed 200 people and had an annual payroll of $7.2 million.

Moore *et al.* (forthcoming) estimated the following impacts of the ETP training program on the employees of firms receiving ETP, as shown in Table 11.9. Using a difference-in-differences approach, they compared the growth of the number of employees, wages paid, and earnings/employee of ETP firms with

Table 11.9. **Growth rates of selected outcomes of workers in firms receiving California ETP training relative to those in a comparison group of firms**

Outcomes		Growth rates	Difference-in-differences
Employees	ETP	14.3%	
	Comparison	–0.8%	15.1%
Wages paid	ETP	25.8%	
	Comparison	10.2%	15.6%
Earnings/Employee	ETP	11.9%	
	Comparison	11.7%	0.2%

Source: Moore, Blake, Phillips, and McConaughy (forthcoming).

those of comparison firms. They found that ETP firms experienced a significantly greater growth in the number of employees and wages paid than did the comparison firms, but they found no significant difference in the growth of earnings per employee.

The evaluators also estimated the impact of ETP on the California economy. Since the goal of ETP was to reduce UI payments and to promote economic development, they examined the impact of ETP on UI savings and earnings. The impact of ETP training is the difference between what occurred with ETP's training programs in place and what would have occurred without them. They estimated that the $73 million invested in ETP resulted in $63 million in UI savings and generated $200 million in additional earnings. Combining these effects and adding indirect effects of $150 million brought the total impact of ETP to over $400 million. The authors caution, however, that the ratio of total impacts to initial investment should not be interpreted as a rate of return of the program, since they did not include all costs and returns in their calculations. For example, they did not include all training costs borne by individual trainees and companies and ETP administrative costs. They also excluded benefits that accrued after the first post-training year, and they ignored benefits to companies beyond what accrued to their employees.

Using a representative firm approach

The representative firm approach used a different methodology to consider the relative importance to a firm of state-financed customized training. This approach does not evaluate the actual behavior of firms, but rather examines the effects of training on the income stream of representative firms. The representative firm approach has been used by several researchers, including Fisher and Peters (1998) and studies by major consulting firms, to analyze the relative merits of business locations. This methodology is based on the income statements and other financial measures of representative

firms in selected industries. The analysis takes into account the benefits to the firm, typically measured with respect to a firm's net present value of future income (NPV) or its internal rate of return (IRR) when the dollar value of the different economic development incentives and subsidies are included. Thus, a firm that enjoys property-tax abatements by locating in a specific area will exhibit cost savings over an identical firm located in another area that does not offer this incentive.

Workforce development activities are entered into this analysis by including only the cost savings of having the government sponsor the training instead of the firm incurring the cost. Where applicable, it also includes the state's reimbursement for lost wages while the worker is in training. Only programs that provide training specifically to businesses to meet their needs are included; federal programs that focus on benefits to individuals are excluded. No attempt is made to estimate the increased productivity of the worker and its effect on the overall productivity of the firm resulting from the training.

Fisher and Peters (1998) provide the most explicit estimates of the various economic development activities using this method. They find that, on average, non-tax incentives are a major part of the state's entire incentive package. In some cases, as much as 90 per cent of state incentives were derived from non-tax programs. Of the three non-tax incentives – workforce training, infrastructure, and general use – workforce training incentives are more important than infrastructure incentives for all size firms in all industries included. Training incentives are the leading non-tax incentive for large firms, defined by number of employees.[20] Not surprisingly, training is not as important as general-use incentives (which include a variety of loan, loan guarantee and subsidy programs primarily to SMEs) to small firms. Overall, Fisher and Peters estimate that the average package of non-tax financial incentives, expressed in terms of present value wage equivalence, was about 9 cents per hour per employee. This means that a typical plant could absorb higher wages by as much as 9 cents an hour for all employees over a 20-year period without affecting its bottom line. For some firms, workforce training accounts for up to half of the savings in payroll. Adding to this analysis the enhanced productivity effects found in other studies would raise the financial advantage of firms who receive these programs.

Evidence of the effectiveness of state-financed training programs

Obviously, without a sufficient number of reliable studies of state-financed job training programs to draw upon, evidence of their effectiveness is spotty. The small group of evaluations of state training programs yields the following conclusions. Holzer, *et al.* (1993) shows that subsidies to private training increase training and improve productivity. Piecing together the essential components that link training to firm productivity and using

estimates from private-sector training suggests that training can increase the cost advantage of firms by raising productivity more than it raised wages in a firm, thus lowering unit costs (Barron, Berger and Black, 1997). However, lower operating costs are not enough to increase employment in an area. It must also be demonstrated that firms respond sufficiently to lower costs by hiring additional workers or by moving to an area that has this cost advantage. The response of firms to these conditions depends upon several factors, including the availability of other factors and the general business climate of the region. Moore *et al.* (forthcoming) find evidence that state training programs increase productivity for both firms and workers, increases wages, improves the work environment, and expands the workforce. These results are based on a handful of studies of specific state training programs. Studies of additional state training programs are required before one can have the same confidence in the robustness of these results as one has in the results from evaluations of federal training programs.

Conclusion

Evaluations have become an integral part of job training programs at the federal level, but to a lesser extent at the state and local levels. Evaluations of federal programs have proven useful to policy makers in determining the reauthorization and design of training programs. The few evaluations that have been conducted of state training programs have also yielded interesting and useful results.

The contrast between evaluations at the federal level and at the state and local level offers insight into the motivation to conduct evaluations and to use their results. The relatively small number of state and local evaluations of training programs, or other social programs, may be explained by three factors: 1) expectations regarding the validity and usefulness of the evaluation results, 2) advocacy and political will, and 3) resource constraints. First, social programs that are intended to affect local and regional economies are difficult to evaluate. Their scope is much broader than the outcomes of individuals, and the scale of their intervention is typically too small to make a significant difference in the region. Therefore, expectations regarding the usefulness of evaluating such programs typically do not outweigh the costs of administering it.

Second, state and local programs typically lack strong advocates for evaluations. Because of the imprecise estimates and the likelihood that the results will show little, if not negative, impacts of the program, groups interested in the program are often reluctant to push for an evaluation. Taxpayers and government watchdog groups tend to pay more attention to financial accounting improprieties than to the ineffectiveness of a program. State and local government officials in charge of economic development

programs are faced with the political and economic reality that they are in competition with other regions to retain and attract businesses. For them, the issue of whether or not to pursue an economic development activity is related more to competition with other jurisdictions than to the effectiveness of the program. These officials often hold the view that if businesses expect to receive economic development incentives, of which customized job training is one, then states and local governments must provide the appropriate incentive if they expect to compete. Thus, in this competitive environment, political will is driven more by perceptions and expectations than by actual effects.

In contrast, the federal government is not subject to the same peer pressure as state and local governments, and it is assured that the evaluations of their programs will generally yield more precise and thus useful results. Furthermore, evaluations provide a means by which politicians from opposing parties can strike a compromise in passing social legislation.

Third, resources can also become a factor, especially for smaller states and local jurisdictions. Resources include not only out-of-pocket expenses to conduct evaluations, but also sufficient expertise and talent on the part of the staff of these jurisdictions who can understand and promote the use of evaluations. The cost of evaluating a program should not be a major barrier to conducting evaluations, however, except for the smaller jurisdictions. Typically, as much if not more money is spent on auditing the financial statements of a program than it would take to evaluate the effectiveness of the program itself. Yet, those interested in assuring that government funds are spent as intended should be as interested as much in the program's effectiveness as they are with its financial transactions. Thus, an important resource consideration is to assure that staff and policymakers have the training and relevant information to appreciate the value of an evaluation and to understand and use its results.

Nonetheless, it should be noted that evaluations were an integral part of many state and local welfare reform initiatives during the 1990s. The difference between those programs and many of the state-financed job training programs is federal involvement. In many cases with welfare reform, states were granted waivers to change various aspects of the federal program only if they agreed to evaluate the program. Much of what we know about the effectiveness of different experiments with welfare reform comes directly from these evaluations.

Therefore, from the examples of evaluating federal job training programs, it is apparent that evaluations are valuable instruments for shaping social policy. It is also evident that higher levels of government, preferably the federal government, need to take the lead in promoting the use of evaluations of social programs. They have the resources and the expertise, and they are

not subject to pressures that diminish the political will to pursue them. Yet, states and local jurisdictions may become more interested in conducting evaluations if it can be demonstrated that evaluations are credible and that they can aid in the continual improvement of their programs. With tight government budgets and taxpayers' impatience with inefficient use of taxpayers' dollars, governments should welcome opportunities to become more cost effective. Evaluations provide valuable information to guide policy and to design programs to better meet the needs of those they are intended to serve.

Notes

1. These comparisons abstract from the Earned Income Tax Credit (EITC) paid to low income workers with dependent children in the US. In recent years the EITC, which is essentially a targeted wage subsidy, totaled about $30 billion or roughly equal to the total expenditures for LMPs listed in the text.
2. When the pre-post estimator of net impacts is applied, the pre-program status of the participant group constitutes the counterfactual.
3. While net impact estimates based on classically designed experiments involving random assignment are considered the "gold standard" of evaluations, they have limitations in the scope of the analysis. Friedman, Greenberg and Robins (1997) point out that net impact evaluations indicate only whether a particular program works, on average, for a particular sample of under a particular set of circumstances. This approach does not open the "black box" to reveal what aspects of the program contributed to its success or failure.
4. Such a selection exercise may be performed using a standardized measure such as the Mahalanobis distance (O'Leary, Nesporova and Samorodov 2001, p. 143).
5. For examples of employment programs evaluated using a classically designed field experiment see Decker and O'Leary (1995).
6. In this report, since the main dependent variable of interest – placement in a normal job – is binary, the regression model predicts the probability of re-employment. The ordinary least squares (OLS) estimation is a linear probability model, which may yield biased estimates. OLS estimates may be biased since the range of variation in the dependent variable is constrained to the zero-one interval. Maddala (1983, Chapter 1) suggests other estimators in such cases. Bias is usually most severe when the bulk of probability clusters at one or other extreme of the zero-one interval. Since re-employment probabilities for the job training and comparison groups generally range from about 40 to 60 per cent, the limited range of the dependent variable is not a likely source of severe bias in estimating parameters by OLS.
7. In this application the regression model is a statement of an analysis of covariance methodology, where X_1 to X_n are the covariates. Mohr (1992, pp. 83-87) discusses extending a regression model for program impacts to include control variables.
8. A method for creating a matched pairs comparison group based on observable characteristics is explained in Annex 4, Section A.4.2.
9. Heckman and Smith (1996, p. 74-78) explain the difference-in-difference method as a fixed effect estimator.

10. A non-parametric approach which also adjusts for unobservable factors is used by Puhani (2002) who provides an example of creating a matched pairs comparison group using participation propensity scores which implicitly account for unobservable as well as observable factors. Impact estimates are then a simple difference of means between the two groups. Heckman and Smith (1996, pp. 72-74) explain the theoretical appeal of the propensity score approach.
11. Technical details of the estimator properties: unbiasedness, efficiency, and robustness are explained in Annex 5 to this manual.
12. These are standard errors of point estimates of program net impacts.
13. An overview of this problem is given by Greene (1993, pp. 273-277).
14. *www.mathematica-mpr.com/3rdLevel/earlyimpact.htm*.
15. A short write-up of this evaluation can be found in *Employment Research*, W.E. Upjohn Institute for Employment Research, October 2002, Vol. 9, No.4 at *www.upjohn.org*.
16. There have been several good recent surveys addressing the question of what type of job training works best for particular demographic groups (King [2002], Barnow and King [2000], Heckman, LaLonde and Smith [1999]).
17. US General Accounting Office, June 1989, p. 33.
18. There are obviously many other programs run by local and state governments and non-government organisations, some of which have been evaluated. Many of these programs are described in Giloth (1998) and Bartik (2001). We focus on state-financed job training programs that target firms in order to focus on programs that are intended to affect local economies and to contrast these with the federal job training programs that target individual workers.
19. For instance, Walker and Greenstreet (1989) find that site selection decisions are sensitive to job training programs.
20. Three plant size categories were used. Large plants were classified as those with employees at or above the 75th per centile of their respective industries. Small plants were those whose number of employers was below the 25th per centile. Therefore, there was no absolute firm size delineation since firm size various significantly across industries.

References

ASHENFELTER, Orley (1978), "Estimating the Effect of Training Programs on Earnings", *Review of Economics and Statistics,* 6(1): 47-57.

BARNOW, Burt S. (1989), "Government Training as a Means of Reducing Unemployment" in *Rethinking Employment Policy,* D. Lee Bawden and Felicity Skidmore, eds. Washington, DC: Urban Institute.

BARNOW, Burt S. and Christopher A. KING (eds), (2000), *Improving the Odds: Increasing the Effectiveness of Publicly Funded Training*. Washington, DC: The Urban Institute Press.

BARNOW, Burt S. and Jeffrey A. SMITH (2002), "Performance Management of US Job Training Programs", presented at the conference *Job Training in the United States: History, Effectiveness and Prospects,* Augusta, Michigan: W.E. Upjohn Institute for Employment Research.

BARRON, John M., Mark C. BERGER and Dan A. BLACK (1997), *On-the-Job Training*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

BARTIK, Timothy J. (1991), *Who Benefits from State and Local Economic Development Policies?* Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

BARTIK, Timothy J. (2001), *Jobs for the Poor: Can Labor Demand Policies Help?* New York, NY: Russell Sage Foundation and MI: W. E. Upjohn Institute for Employment Research.

BURGHARDT, John, Peter Z. SCHOCHET, Sheena MCCONNELL, Terry JOHNSON, R. Mark GRITZ, Steven GLAZERMAN, John HOMRIGHAUSEN and RusselL JACKSON (2001), *Does Job Corps Work? Summary of the National Job Corps Study.* Document No. PR01-50. Princeton, NJ: Mathematica Policy Research, Inc. (June).

DECKER, Paul T. and Christopher J. O'LEARY (1995), "Evaluating pooled evidence from the re-employment bonus experiments". *Journal of Human Resources*, Vol. 30, No. 3, pp. 534-550.

DECKER, Paul T. and Irma PEREZ-JOHNSON (2002), "Individual Training Accounts and Eligible Provider Lists", presented at the conference *Job Training in the United States: History, Effectiveness and Prospects,* Augusta, Michigan: W.E. Upjohn Institute for Employment Research.

FISHER, Peter S. and Alan H. PETERS (1998), *Industrial Incentives: Competition among American States and Cities*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

FRIEDLANDER, Daniel, David H. GREENBERG and Philip K. ROBINS (1997), "Evaluating Government Training Programs for the Economically Disadvantaged", *Journal of Economic Literature* 35(4), December.

GILOTH, Robert (1998), *Jobs and Economic Development: Strategies and Practice*. Thousand Oaks, CA: Sage Publications.

GREENE, William H., (1993), *Econometric Analysis*. New York: Macmillan.

HECKMAN, James J., (1979), "Sample Selection Bias as a Specification Error". *Econometrica* 47(1): 153-161.

HECKMAN, James J., Robert J. LALONDE and Jeffrey A. SMITH (1999), "The Economics and Econometrics of Active Labor Market Programs", in *Handbook of Labor Economics, Volume 3A*, Orley Ashenfelter and David Card, editors. Amsterdam: Elsevier.

HECKMAN, James, REBECCA Roselius and Jeffrey SMITH (1994), "US Education and Training Policy: A Re-evaluation of the Underlying Assumptions Behind the 'New Consensus'". In *Labor Markets, Employment Policy, and Job Creation*, Lewis Solmon and Alec Levenson, eds. Boulder, CO: Westview Press.

HECKMAN, James J. and jeffrey A. SMITH (1996), "Experimental and Nonexperimental Evaluation", in Schmid, O'Reilly, and Schomann, eds. *International Handbook of Labour Market Policy and Evaluation*. Cheltenham, UK: Edward Elgar, pp. 37-88.

HOLLENBECK, Kevin M. (2002), "Washington's Workforce Development System Pays Off". *Employment Research*. 9(4): 4-6. Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

HOLZER, Harry J., Richard N. BLOCK, Marcus CHEATHAM and Jack H. KNOTT (1993), "Are Training Subsidies for Firms Effective? The Michigan Experience", Industrial and Labor Relations Review, Vol. 46(4), July, 635-636.

KING, Christopher A. (2002), "The Effectiveness of Publicly Financed Training in the US", presented at the conference *Job Training in the United States: History, Effectiveness and Prospects,* Augusta, Michigan: W.E. Upjohn Institute for Employment Research.

KLUVE, Jochen and Christoph M. SCHMIDT (2002), "Can Training and Employment Subsidies Combat European Unemployment?" *Economic Policy: A European Forum.*

LEIGH, Duane E. (1990), *Does Training Work for Displaced Workers?: A Survey of Existing Evidence*. Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

MADDALA, G.S. (1983), *Limited-dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.

MALLAR, Charles, Stuart KERACHSKY, Craig THORNTON and David LONG (1980), *An Evaluation of the Economic Impact of the Job Corps Program*. Project Report 80-60. Princeton, NJ: Mathematica Policy Research, Inc.

MANGUM, Garth L. (1968), *MDTA: Foundation of Federal Manpower Policy*. Baltimore, MD: The Johns Hopkins Press.

MOHR, Lawrence B. (1992), *Impact Analysis for Program Evaluation*. London: Sage.

MOORE, Richard W., Daniel BLAKE, G. Michael PHILLIPS and Daniel MCCONAUGHY (forthcoming), *Training that Works: Lessons from California's Employment Training Panel Program*. Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

OECD (2001), *OECD Employment Outlook*. Paris: Organization for Economic Cooperation and Development (June).

O'LEARY, Christopher J., Alena NESPOROVA and Alexander SAMORODOV (2001), *Manual on Evaluation of Labour Market Policies in Transition Economies*. Geneva: ILO.

O'LEARY, Christopher J. and Robert A. STRAITS (2002), "Intergovernmental Relations in Employment Policy: The United States Experience", in *Labour Market Policy and Federalism: A Comparative Perspective*, Alain Noël, ed. Montreal and Kingston: McGill-Queen's University Press.

ORR, Larry L., Howard S. BLOOM, Stephen H. BELL, Fred DOOLITTLE, Winston LIN and George CAVE (1996), *Does Training for the Disadvantaged Work?: Evidence from the National JTPA Study.* An Abt Associates Study. Washington, DC: Urban Institute Press.

PUHANI, Patrick A. (2002), "Advantage through Training in Poland?: A Microeconometric Evaluation of the Employment Effects of Training and Job Subsidy Programs". *Labour* 16(3): 569-608.

SOCIAL POLICY RESEARCH ASSOCIATES (2001), *PY 99 SPIR Data Book*. Oakland, California: Social Policy Research Associates (April).

STROMSDORFER, Ernst W. and George FARKAS (eds) (1980), *Evaluation Studies Review Annual – Volume 5*. Beverly Hills, CA: Sage.

US GENERAL ACCOUNTING OFFICE (1989), "Job Training Partnership Act: Services and Outcomes for Participants with Different Needs", US Government Printing Office, June, p. 33.

WALKER, Robert and David GREENSTREET (1989), "Public Policy and Job Growth in Manufacturing: An Analysis of Incentives and Assistance Programs". Paper presented at the 36th North American Meeting of the Regional Science Association, Santa Barbara, CA, November 10-12.

WANDNER, Stephen A., David E. BALDUCCHI and Amanda B. SPICKARD (2001), "Expenditures on Active and Passive Labor Market Policy in the United States Estimates for Fiscal Year 2001". Presented at the international workshop "Active Labor Market Programs: Improvement of Effectiveness" sponsored by Ministry of Labor and Social Development of the Russian Federation, World Bank, and Russian Foundation for Social Reforms, October 2-3, Moscow, Russia.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 12

Evaluating Local Economic Development Policies: Theory and Practice*

by
Jeffrey Smith,
Department of Economics, University of Maryland, USA

* This chapter has benefited from discussions with and/or comments from Tim Bartik, Dan Black, Michael Lechner, Alistair Nolan, Miana Plesca, Elliot Stern and Alex Whalley, from reading the chapters by Tim Bartik and by Randy Eberts and Chris O'Leary (which were written before this one) and from discussions at the OECD conference in Vienna in November 2002.

1. Introduction

Policies and programs undertaken to increase local economic development by governments and by private agencies may have positive effects, or they may not. In some cases, a lack of effects may result from poor program design or inadequate funding. In other cases, a lack of effect may result from the fact that the program really exists to funnel money to politically influential firms, individuals or groups, with the local economic development justification used as cover. When programs do not produce benefits in terms of local economic development, finding this out allows scarce funds to flow into other, more beneficial activities, or back to the longsuffering taxpayer. When programs do produce benefits, finding this out can generate political support for program persistence or even expansion.

Evidence on the efficacy of local economic development policies and programs comes from evaluations. This chapter presents an overview of the current literature on how to evaluate programs. The scholarly literature on program evaluation has advanced rapidly over the past fifteen years. For example, major developments in regard to "heterogeneous treatment effects" – different program impacts for different persons, firms, counties, cities or groups affected by a policy – affect both evaluation practice and how to think about evaluation design and interpretation. Similarly, important technical developments in non-parametric and semi-parametric methods allow much more flexible use of the available data, but at the same time create a demand for the high quality data that such methods require to produce reliable estimates. Social experiments have become routine (at least in the United States) in areas such as the evaluation of public employment and training programs. Unfortunately, evaluation practice, to a large extent, remains mired in the 1970s. One of the main goals of this chapter is to provide a practical, relatively non-technical guide to these advances.

This chapter addresses some of the same issues as the chapters by Tim Bartik (2004) and by Randy Eberts and Chris O'Leary (2004), but with enough differences to make it a complement to, rather than a substitute for, those chapters. Five differences in particular deserve notice. First, this chapter devotes much more attention to the different econometric evaluation estimators in the literature, and provides a wealth of pointers into the rapidly expanding literature on the subject. A key theme of the chapter is the choice of an appropriate estimator given the available data, economic environment

and institutional characteristics of the program being evaluated. Second, this chapter devotes more attention to the emerging literature on heterogeneous treatment effects, and how such effects influence evaluation design and interpretation. Third, this chapter worries more about the implications of general equilibrium effects for policy evaluations. Fourth, this chapter emphasizes that doing an evaluation may not make sense in all cases, particularly for smaller programs. Time spent reading the literature for good evaluations of similar programs may yield more useful results than a weak evaluation based on poor data completed by a poorly qualified evaluator using inappropriate methods. Finally, the perspective underlying this chapter is that evaluation, taken seriously, represents a method for ensuring that program managers further the goals of their principals – namely taxpayers and donors – rather than simply transferring resources to interested stakeholders, such as program operators, politically favoured firms, or themselves. In practice, many low quality evaluations exist mainly to cover up exactly such behaviour; for precisely this reason it is important to be very clear about what constitutes a good evaluation and to design institutions that will reduce the flow of misleading, low-quality evaluations.

The remainder of the chapter is organized as follows. Section 2 describes the evaluation problem and discusses parameters of interest. Section 3 provides an overview of the theory of econometric program evaluation at a relatively non-technical level, and with plenty of pointers to the literature. Section 4 reviews the two leading serious alternatives to econometric program evaluation: participant self-evaluation and administrative performance standards. Section 5 discusses the practice of evaluation, in the broad sense of the choices facing an organisation considering undertaking an evaluation, such as whether or not it is worth it to do an evaluation, who should do the evaluation, and how to make sure that it is any good. Section 6 concludes and restates the main themes of the chapter.

2. Programs and parameters

2.1. Types of local economic development programs

Local economic development programs include a wide range of initiatives, from programs designed to improve the human capital of individual workers, to financial and in-kind subsidies to professional athletic teams, to enterprise zones, to tax subsidies designed to lure particular businesses, and on and on. Bartik (2004) presents a nice list in his Table 12.1*a* that includes a somewhat narrower set of activities than I have in mind here; Bartik (2003a) describes these policies in greater detail.

For this chapter, two dimensions of such programs hold particular relevance, as they shape choices regarding data collection and evaluation

methods, as I describe in detail below. The first dimension consists of the units directly treated by the intervention. Depending on the program, this could be individual workers, some or all firms in an area, cities, towns or districts, or entire states or countries. The second, related, dimension consists of the units that theory suggests the program will affect. In some cases, particularly for programs not expected to have much in the way of external effects, these two dimensions may coincide. For example, small-scale human capital programs may have little effect on individuals other than those receiving the additional human capital. In other cases, the two dimensions will not coincide. For example, a program may have positive effects on treated units and negative effects on untreated units, as when subsidizing one class of firms but not their competitors. In still other cases, programs may produce positive spillovers, as when a new park attracts new businesses and residents to an area, and increases property values in the surrounding neighborhoods.

2.2. Notation

Popular and policy discussions of economic development programs often focus on their "effects," as though the "effects" of a program represent a single well-defined entity. That programs have a variety of effects represents an important theme of this chapter. In the academic literature, this discussion falls under the heading of heterogeneous treatment effects. That literature discusses how the notion of a program's effects changes and broadens when we consider that a program may have a different effect on each unit that participates in it and, in some cases, even on units that do not participate in it.

To make this point more clearly, I now introduce some very simple notation, which will serve to make meanings precise throughout the chapter. However, the chapter is written so that it does not require an understanding of the notation to get the point; severely notation-averse readers can simply skim over it.[1]

Let *Y* denote some outcome variable. For an individual, it might be earnings, employment, or health status. For a firm it might be profits, sales, or employment. For a locality it might be population, or some measure of air quality or economic growth. Now imagine two worlds for each unit, one world where the unit participates in the program under study and one where it does not. We can imagine the unit's value of *Y* in each of those worlds, and we label the value in the world where the unit participates as Y_{1i} and the value in the world where the unit does not participate as Y_{0i}, where "i" refers to a particular unit.

2.3. Parameters of interest

Using this notation, the effect of a program on unit "i" is given by

$\Delta i = Y_{1i} - Y_{0i}$.

In words, the literature defines the effect of a program on unit "i" as the difference in outcomes between a world where that unit participates and a unit where it does not. The evaluation problem then consists of estimating whichever one of the two outcomes we do not observe in the data.

Many of the various parameters of interest defined and examined in the literature on program evaluation then consist of averages of the unit-specific impact (Δ_i) over various policy-relevant sets of units. The most common parameter of interest is the Average Treatment Effect on the Treated (ATET), or just "treatment on the treated" for short. This parameter indicates the average effect of the program on current participants. In terms of the notation just defined, it equals

$$\Delta^{TT} = E(D_i \mid D_\iota = 1) = E(Y_{1\iota} - Y_{0i} \mid D_\iota = 1),$$

where D_i is a dummy variable for current participants, so that $D_i = 1$ for units that participate in the program and $D_0 = 0$ otherwise, and E denotes the expectations operator, where the expectations are conditional on the condition to the right of the vertical bar ("|"). An estimate of the ATET, combined with an estimate of the average cost of the program per participating unit, allows a cost-benefit analysis of the question of whether to keep or scrap an existing program.

The Average Treatment Effect (ATE) represents a second parameter of potential interest. This parameter averages the effect of treatment over all of the units in a population, including both participants and non-participants. In terms of the notation,

$$\Delta^{ATE} = E(Y_{1\iota} - Y_{0i}).$$

The ATE answers policy questions related to universal programs – programs where every unit in some well-defined population participates. When considering making a voluntary program mandatory, policymakers need precise estimates of the ATET and the ATE, which may differ strongly if those units that choose not to participate in a voluntary program do so because it would have only a small, or even negative, effect for them. An example here would be taking a voluntary program of job search assistance or job training for displaced workers and making it mandatory (or nearly so by, for example, requiring participation in order to receive full unemployment insurance benefits).

A third category of parameters consists of Marginal Average Treatment Effects, or MATEs. A marginal average treatment effect measures the average effect of a program among a group at some relevant margin. For example, suppose that the program under consideration presently serves firms with fewer than 20 employees, and the proposal under consideration consists of expanding it to include firms with 21 to 30 employees. In this situation, the parameter of interest consists of the average effect of participation on firms with 21 to 30 employees. This parameter may differ from the ATET, which

would give the average effect on existing participant firms (those with 1 to 20 employees), and could be either higher or lower, depending on the nature of the program treatment and its relationship to firm size. Comparing a MATE to the corresponding marginal cost of expanding (or contracting) a program provides a cost-benefit analysis for the program expansion or contraction. Note that a different MATE applies to each margin – expanding a program to include one set of units may yield different results than expanding it to include another set of units. A final category of parameters, called Local Average Treatment Effects, or LATEs, is discussed in Section 3.4.

The parameters presented so far may or may not capture general equilibrium effects, depending on the design of the analysis. General equilibrium effects are those other than the immediate effects on the treated units, and result from changes in the behavior of untreated units in response to the program. Such changes may occur directly (a firm with 11 employees fires one in order to become eligible for a program that serves firms with 10 or fewer employees) or indirectly through changes in prices, as when a tuition subsidy increases the supply of skilled workers and thereby lowers their wage. Consider a state-level program that subsidizes training at a particular class of firms. Some states have the program and others do not. An estimate of the ATET on the firms receiving the subsidy will capture only the direct effects on the employment, productivity, sales, and so on for those firms. In contrast, an estimate of the ATET on the states adopting the subsidy will capture any general equilibrium effects at the state level, including reductions in employment at unsubsidized firms, but not general equilibrium effects that operate across state boundaries.

Bringing general equilibrium effects into the picture adds some additional parameters of interest. For example, we might now have some interest in what the literature calls the Average Effect of Treatment on the Non-Treated (ATNT). This parameter measures the average effect of the program on units that do not participate in it, either because they choose not to or because they do not meet the eligibility criteria. To see this, consider the case of a program that provides job search assistance to particular groups of workers. These workers now search more, and more intelligently, than before, and we would expect them to find jobs faster. But what happens to others in the labor market? First, some jobs that would have been filled by others now get filled by individuals who receive the job search assistance. As a result, they find jobs more slowly. Second, it may make sense for them to change their search intensity as well. Both factors lead a program that provides services to one group to have effects on other groups – effects that matter in assessing the value of the program. Calmfors (1994) provides a useful (and relatively accessible) introduction to general equilibrium issues in the context of active labor market policies.

3. Theory

This section provides a brief introduction to each of the main categories of econometric evaluation methods. Sections 3.1 to 3.6 each consider one category of method, and Section 3.7 considers how to choose among them. Although this chapter presents the various estimators as though they are dishes on a buffet, where the evaluator can choose which one to use based on its having a cool name, or its association to famous people, or its being the estimator *de jour*, in fact, estimator selection, properly done, must adhere to strict rules. Each of the categories of estimators examined in the following sub-sections provides the correct answer only under certain assumptions. An evaluator choosing an estimator must carefully consider the nature of the available data, the institutional nature of the program – particularly how participation comes about – and the parameters of interest. In some cases – and this constitutes another one of my themes – a lack of good data may mean that no estimator is likely to provide a correct answer, in which case the evaluator should simply stop and report this fact.

For readers wanting to learn more, the literature provides a number of other surveys of all or part of this material, ranging from the very non-technical to the very technical. At the less technical end, see Moffitt (1991, 2003), Winship and Morgan (1999), Smith (2000), Ravallion (2001), and Smith and Sweetman (2001). At a moderate technical level, see Angrist and Krueger (1999), Blundell and Costa Dias (2000,2002), and Heckman, LaLonde and Smith (1999), except for Section 7. For strongly technical presentations see Section 7 of Heckman, LaLonde and Smith (1999) and Heckman and Vytlacil (2004). Some standard econometrics texts contain presentations that emphasize the issues focused on in the evaluation literature. In this regard, see Wooldridge (2002) at the undergraduate level and Green (2002) or Wooldridge (2001) at the graduate level.

3.1. Social experiments

Social experiments represent the most powerful tool in the evaluator's toolbox, but just as that favorite wrench may not make a good screwdriver, so social experiments serve the evaluator better in some contexts than in others. To see why evaluators like social experiments, consider a treatment with no external effects, and suppose that we seek to determine the impact of treatment on the treated. The primary problem in evaluation research (almost) always consists of non-random selection into treatment. Because of non-random selection into treatment, one cannot simply compare the outcomes of treated units with the outcomes of untreated units in order to determine the impact of treatment. In terms of the notation defined above, we cannot rely on the average outcomes of untreated units, $E(Y_0 \mid D = 0)$, to

accurately proxy for the outcomes that treated units would have experienced, had they not been treated, $E(Y_0 \mid D = 1)$. Finding a good approximation to this counterfactual represents the tough part of estimating the treatment on the treated parameter (because the outcomes of participants, $E(Y_1 \mid D = 1)$, appear directly in the data). The problem of non-random selection into treatment is called the selection bias problem in the econometric evaluation literature. It is important to distinguish the classical selection bias problem of selection on the untreated outcome with non-random selection into treatment based on the effect of treatment. This latter type of selection has only recently received substantial attention in the literature; this type of selection is what makes, for example, the mean impact of treatment on the treated different from the average treatment effect in programs that do not treat all eligible units.

Social experiments solve the selection bias problem by directly constructing the usually unobserved counterfactual of what participating units would have experienced, had they not participated. In particular, in a social experiment, units that would otherwise have received the treatment are randomly excluded from doing so. The outcomes of these randomly excluded units, under certain assumptions, provide an estimate of the missing counterfactual mean, given by $E(Y_0 \mid D = 0)$. This ability to obtain the counterfactual under what, in many (but not all) contexts represent very plausible assumptions, defines the power of experiments, and explains their attraction to evaluators.

As the virtues of experiments are fairly well known, and also extensively detailed in Bartik's chapter, I focus instead on some of the conceptual issues and limitations associated with experiments. The purport of this discussion is not to provide cover to those who want to avoid doing experiments because they wish to maintain an aura of uncertainty about the impacts of the programs they love (or benefit from financially, which often amounts to the same thing). Rather, it is to make it so that experiments do not get used when they do not or cannot answer the question of interest, and to make sure that they get interpreted correctly when they are used.

The first limitation of experiments is that they cannot answer all questions of interest. This limitation has three facets, which I cover in turn. First, randomization is simply not feasible in many cases. The evidence suggests that democracy increases economic growth, but we cannot randomly assign democracy to countries. Similarly, political factors may prohibit randomization of subsidies to firms or randomization of development grants to cities and towns.

Second, experimental data may or may not capture the general equilibrium effects of programs. Whether or not they do depends on the units affected by any equilibrium effects and the units that get randomized in the experiment. If there are spillovers to units not randomized, as when a program

for small firms has an effect on medium-sized firms, these effects will be missed. Similarly, positive spillovers will get missed in an evaluation that randomizes only treated firms, rather than randomizing at the locality level.

Finally, experiments provide the distribution of outcomes experienced by the treated, and the distribution of outcomes experienced by the untreated. They do not provide the link between these two distributions; put differently, experimental data do not indicate whether a treated unit that experienced a very good outcome would also receive a very good outcome had it not received treatment. In technical terms, an experiment provides the marginal outcome distributions but not the joint distribution. As a consequence, without further, non-experimental, assumptions, experimental data do not identify parameters that depend on the joint distribution of outcomes, such as the variance of the impacts. See Heckman, Smith and Clements (1997) for an extended discussion of this issue and a variety of methods for obtaining estimates of these parameters.

The second major limitation of experiments is that practical difficulties associated with the implementation of the experiments can sometimes complicate their interpretation. Readers interested in more general treatments of the implementation of social experiments should consult, *e.g.,* Orr (1998), as well as the implementation reports or summaries associated with major experimental evaluations, such as Hollister, Kemper and Maynard (1984), Doolittle and Traeger (1990), Newhouse (1994) and so on. The *Digest of the Social Experiments*, compiled by Greenberg and Shroder (1997), presents a comprehensive list of all the social experiments, along with pointers to details about their design, implementation and findings.

First, because an experimental evaluation tends to have a greater disruptive effect on local program operation than a non-experimental evaluation, experiments in decentralized or federal systems, such as those in the US and Canada, often have problems with external validity, because of non-random selection of local programs into the experiment. This was an issue in the US National JTPA study, where over 200 of the 600 local training centers were approached in order to find 16 willing to participate in the experimental evaluation. Other than trying to keep the experimental design relatively unobtrusive, and offering side payments (about US$1 million was devoted to this in the JTPA Study), little can be done about this other than comparing the characteristics of participating and non-participating local programs and avoiding overly ambitious generalizations about the results.

Second, as described in, *e.g.,* Heckman and Smith (1995), experiments may suffer from randomization bias. This occurs when individuals behave differently due to the presence of randomization. For example, if the units under study can undertake activities that complement the treatment prior to

receiving it, they have less incentive to do so during an experiment, because they may be randomly excluded from the treatment. Note that randomization bias differs from Hawthorne effects. The latter occur when individuals being evaluated change their behavior in response to being observed, whether in the context of an experimental or a non-experimental evaluation. Little empirical evidence exists on the importance of randomization bias.

Third, depending in part on the placement of random assignment within the process by which units come to receive the treatment, dropout within the treatment group may cause problems for the interpretation of the experimental impact estimates. Dropout here refers to a departure from the treatment after random assignment, perhaps because it appears less attractive once fully known. Randomly assigning units early in the participation process tends to increase dropout. As detailed in Heckman, Smith and Taber (1998), dropout is a common feature of experimental evaluations of active labor market policies. The usual responses take two forms. In the first, the interpretation of the impact estimate changes and becomes the mean impact of the offer of treatment, rather than of the receipt of treatment. In the second, the impact estimate gets adjusted using the method of Bloom (1984). See Heckman, LaLonde and Smith (1999), Section 5.2, for more details and a discussion of the origins of the adjustment.

Fourth, as discussed in detail in Heckman, Hohmann, Smith and Khoo (2000), in some contexts, control group units may receive a treatment similar to that offered the experimental treatment group from other sources. Their analysis considers the case of employment and training programs, and they show that, at least in the decentralized US institutional environment, where many federal and state agencies offer subsidized training of various sorts, substitution is quite common. In this case, the outcomes of the control group do not represent what the treated units would have experienced had they not received treatment. Instead, they represent some combination of untreated and alternatively treated outcomes. The literature indicates three responses to substitution bias. As with dropouts, one consists of reinterpretation of the parameter – this time as the mean difference between the treatment being evaluated and what the treated units would have received were that treatment not available, which will sometimes be no treatment and sometimes be some other treatment. The second response consists of adjusting the experimental mean difference estimate by dividing it by the difference in the fraction treated between the treatment and control groups. This represents a generalization of the Bloom (1984) estimator and requires that the substitute treatment have a similar impact to the treatment being evaluated. Finally, the third response consists of using the experimental data to do a non-experimental evaluation. See Heckman, Hohmann, Smith and Khoo (2000) for more details.

As discussed in, *e.g.*, Smith and Sweetman (2002), variants of random assignment can sometimes overcome the political obstacles to experimental evaluation. One such design is the so-called "randomized encouragement" design, which applies to voluntary programs with participation rates less than 100 per cent. Here, rather than randomizing treatment, an incentive to participate, such as an additional subsidy or additional information about the program, gets randomly assigned to eligible units. In simple terms, this strategy creates a good instrument (see Section 3.4 for more about instruments) by creating some random variation in participation. This method depends crucially on having an incentive that actually does measurably affect the probability of participating. The second alternative design consists of random assignment at the margin, as in Black, Smith, Berger and Noel (2003). They examine the effect of mandatory reemployment services on unemployment insurance recipients in the state of Kentucky. Individuals get assigned to the mandatory services based on the predicted duration of their unemployment spell. Only individuals at the margin of getting treated get randomly assigned. This proved much less intrusive than full-scale random assignment and also satisfied the state's concerns about treating all claimants with long expected durations. In a heterogeneous treatment effects world, neither of these alternative versions of random assignment estimates the mean impact of treatment on the treated. However, the parameters they do estimate may have great policy interest, if policy concern centers on marginal expansions or contractions of the program.

In sum, experiments have enormous power, both because of their statistical properties and, not unrelated, because of their rhetorical properties. Policymakers, pundits and plebes can all understand experimental designs, something that is not so true of non-experimental methods such as matching, instrumental variables or structural general equilibrium models. In contrast to the present situation, where evaluators must constantly cajole and prod resistant agencies to undertake random assignment evaluations, in a well-ordered polity, government officials would bear the burden of making a case for not doing random assignment in the case of expensive or important programs that justify a full-scale evaluation and that do not fall into the inappropriate categories described earlier in this section.

3.2. *Selection on observables: regression and matching*

Selection on observables occurs when observed characteristics determine participation in a program but, *conditional on those characteristics*, participation does not depend on outcomes in the absence of participation. In such situations, conditioning on the characteristics that determine participation suffices to solve the selection bias problem. In general, selection on observables has the greatest plausibility when the observed data contain

variables that relate to all of the major factors identified by theory (and evidence on similar programs) as affecting both participation and outcomes. A simple example makes the point clear. Suppose that both men and women choose at random to participate in some training program, but that men choose to participate with a higher probability than women. Assume as well that women have better labor market outcomes without the training than do men (not an unrealistic assumption in the populations targeted by many training programs) and that the training has the same average effect on men and women. Simply comparing the outcomes of all participants to all eligible non-participants will understate the impact of the program, because this comparison conflates the impact of training with the effects of the over-representation of men in the program. Because men have worse labor market outcomes in the absence of training than women, the over-representation of men in the program will make this simple comparison a downward biased estimate of the impact of the program. If, instead, we separately compare male participants and non-participants and female participants and non-participants, we will obtain an unbiased estimate of program impact.

By far the most common way of taking account of selection into treatment on observable characteristics consists of using standard linear regression methods, or their analogs such as logit and probit models for limited dependent variables, and including the observables in the model. A standard formulation would look like

$$Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

where Y_i is the outcome of interest, D_i is a dummy variable for receiving treatment (with β_D the corresponding treatment effect), $X_{1i},\ldots,X_{ki}$ are the confounding variables and where the regression would be estimated on a sample of treated and eligible non-treated units (which means you cannot use this approach for a treatment that reaches all eligible units). In a common effect world, provided the selection on observables assumptions holds, β_D estimates the common treatment effect. In a heterogeneous treatment effects world, it estimates the impact of treatment on the treated under fairly general assumptions.

Regression has the great advantages of familiarity and ease of interpretation and use. All standard statistical packages include it, and even some database programs. The coefficients have interpretations as partial derivatives or finite differences (though this becomes a bit more complicated in logit and probit models, some statistical packages now report marginal effects, which are close cousins to partial derivatives.[2]) Despite these advantages, it is important to note that regression is not, in general, an "expedient and satisfactory net impact technique," as claimed in the Eberts and O'Leary chapter in this volume. Whether or not regression produces consistent

estimates depends on whether the selection on observables assumption holds, which in turn depends on the richness of the set of the variables available for inclusion in the regression and on the nature of the selection process in each context. As I emphasize in Section 3.7 below, no econometric evaluation estimator provides a general solution to the selection bias problem. In some cases regression will do so and in others it will not. One of the primary contributions of the evaluator consists in determining which case corresponds to the evaluation at hand, a task that requires more than soothing phrases.

In addition to assuming selection on observables, standard linear regression analysis also imposes a linear functional form on the data, which may or may not correspond to the underlying population relationship. Including higher order terms in X_i relaxes this constraint, but this is rarely done in practice. Matching methods, which have received a lot of attention in the evaluation literature in the past few years, relax the linear functional form assumption inherent in the standard regression approach while maintaining the assumption of selection on observables.

The basic idea of matching is to directly compare individuals with exactly the same (or similar) values for the relevant confounding variables. This avoids any functional form restrictions. The easiest way to see this is to consider an example with only discrete *X* variables. In that case, the simplest version of the matching estimator consists of finding, for each treated unit, an untreated unit with identical values of the covariates. The impact estimate, which provides an estimate of the impact of treatment on the treated in a heterogeneous effects world, then consists of the mean outcome for the treated units minus the mean outcome for the matched untreated units. An estimate of the average treatment effect can be obtained by re-weighting the X-specific estimates by the distribution of the *X* in the population (rather than implicitly weighting them by the distribution of *X* in the treatment group, as matching normally does).

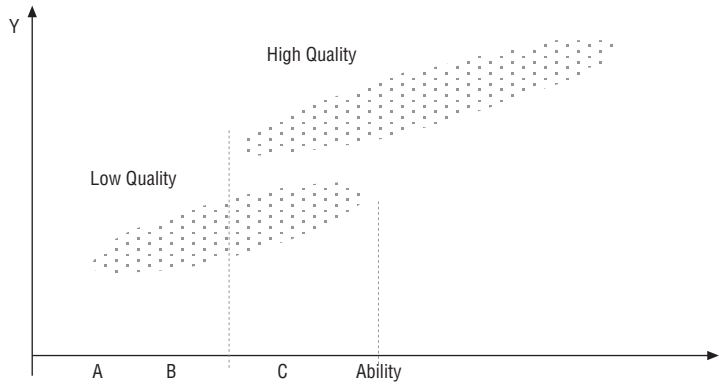
When the set of matching variables includes continuous variables, or a large number of discrete variables, exact matches become difficult.[3] In such cases, matching relies on a distance measure to determine which untreated observations should play a role in estimating the counterfactual for each treated observation. Put differently, the distance metric converts distances on a vector of variables into a single number, which can then be used to match similar observations. With a single continuous variable, absolute differences in that variable can serve as the metric. With a richer covariate set, the choices include Mahalanobis metric matching, as in Rosenbaum and Rubin (1985), and propensity score matching, as developed by Rosenbaum and Rubin (1983). The propensity score is just statistical jargon for the probability of participation conditional on X.

The economics literature has tended to focus on propensity score matching because theory often provides guidance on which variables should affect the probability of treatment and because the probability of participation is often of independent interest and so would get calculated anyway. With propensity score matching, the estimated counterfactual for a given treated observation is based on the outcomes of untreated observations with similar probabilities of participation. In order to satisfy the assumption of selection on observables, the propensity score model should include all of the important variables that affect both participation and outcomes – not just all of the ones in the data set at hand.

Matching methods have the additional benefit, relative to standard regression methods, of focusing attention on the so-called "support" problem. The support problem arises when the data contain no similar untreated observations for some of the treated observations. In such situations, it is common when applying matching methods to simply drop such treated observations from the analysis, which can substantially affect the nature of the parameter being estimated if there are a lot of them. In contrast, standard regression methods will produce estimates even in the absence of comparison units that look like the treatment units because the linear functional form fills in for the missing data. Put differently, the regression identifies the untreated outcome model in the region of the data where the untreated observations lie, and then projects it out into the region of the data where the treated units lie, thereby implicitly estimating the counterfactual.

Figure 12.1 helps illustrate this point. The horizontal axis represents the matching variable under the assumption that conditioning on this single X takes account of non-random selection into treatment. The vertical axis represents the outcomes. The two clouds of data represent treated and untreated units, respectively. Region B, in the center of the figure, constitutes the region of "common support", where there are both treated and untreated observations with roughly similar values of X (close enough to be "good" matches). The support condition fails in Region C, on the right side of the figure, which includes treated observations but no untreated observations with similar values of X. No matching estimate can be constructed for the treated observations in Region C. In sharp contrast, there would be no problem including them in a regression analysis, whose linear functional form would project the conditional mean outcome without treatment estimated using the data on untreated units in Regions A and B out into Region C. In the end, projections of this sort may work out in a given context, but the evaluator would still like to know whether his or her estimates rely heavily on the linear functional form or not. Finally, Region A contains untreated units but no treated units. When estimating the impact of treatment on the treated, this poses no problem; matching ignores these units, as they are not required to

Figure 12.1.



construct the counterfactual for any of the treated units. Standard linear regression methods, on the other hand, make use of these observations to help pin down the relationship between X and the outcome. See Heckman and Vytlacil (2001a) and Black and Smith (2004) for extended discussions of the support problem. Whether matching on P(X) or on some other distance measure based on X, a variety of different methods exist for actually doing the matching when the data lack exact matches. These methods differ on several dimensions, of which the two most important are probably whether or not they use more than one untreated observation to construct the counterfactual for each treated observation, and whether or not they use each untreated observation to help construct the estimated counterfactual for more than one treated observation. The simplest form of matching – common in the applied statistics literature but not in the economics literature – is single nearest neighbor matching without replacement. In this case, a single untreated observation estimates the counterfactual for each treated observation, and untreated observations can only estimate the counterfactual for one treated observation, even if they are the closest untreated observation to many treated observations. This method has the disadvantages that it tends to throw out a lot of comparison observations and that, in data sets with only a few comparison observations, it tends to lead to bad matches, in the sense of pairing treated observations with untreated observations that do not look much like them. Dehejia and Wahba (1999) provide a useful discussion of this issue, with an empirical example that illustrates the dangers of matching without replacement.

Other versions of matching include kernel matching, nearest neighbor matching with more than one nearest neighbor (this increases the bias but lowers the variance), local linear matching and weighting by the inverse of the

propensity score. See Smith and Todd (2004) for a relatively applied overview of the different methods, and, among others, Heckman, Ichimura and Todd (1997, 1998), Heckman, Ichimura, Smith and Todd (1998), and Hahn (1998) for the technical details. Imbens (2004) provides a readable survey of the recent technical literature while Heckman and Navarro-Lazano (2004) contrast propensity score matching methods with methods based on exclusion restrictions and explicate the differing role the estimated probability of treatment plays in each one. Frölich (2004) presents a Monte Carlo analysis that suggests that kernel matching (and another method called ridge matching) tend to outperform nearest neighbor matching, local linear matching and weighting by the inverse of the estimated propensity score. Lechner (1999, 2000) provides some fine examples of matching in practice.

A number of important extensions to matching exist in the literature. For example, longitudinal data allow the combination of matching with the difference-in-differences methods discussed in Section 3.4. Heckman, Ichimura, Smith and Todd (1998) develop these methods. See Blundell and Costa Dias (2002) for a less technical introduction (that also includes the case of repeated cross-section data) and Eichler and Lechner (2002) for an application. In studies of programs that may create effects at the level of a town or small city, matched local area designs are sometimes undertaken. These designs typically share the problem of having too few treated and untreated areas for reliable statistical inference. See Long and Wissoker (1995) for an example.

In sum, matching represents an important extension of traditional linear regression approaches when the data support an assumption of selection on observables. Matching does not solve the problem of selection on unobservables, but does allow flexibility in conditioning and makes it easy to examine the support issue. Matching, whether semi-parametric (propensity score matching) or non-parametric (cell matching), requires more data than traditional approaches that impose more structure on the problem. While the standard statistical packages do not yet contain routines to perform matching, several user-written routines exist for use with Stata.

3.3. Selection on unobservables: longitudinal methods

One very simple method for evaluating a development policy consists of examining the difference in the outcomes of the units affected by the policy before and after the policy comes into force. The implicit assumption underlying this simple strategy is that units subject to the policy change would have had the same outcomes as before, had the policy not intervened to change them. Though reasonable in some contexts, this assumption requires that treated units not select into treatment based on temporary changes in their outcomes. For example, if firms only choose to participate in a subsidy program when they are having a bad year, and if most bad years are

followed by good years even without the subsidy, then a before-after comparison of the outcomes of participating firms will overstate the impact of the subsidy on firm performance by attributing to it the subsidy firms' normal return to good times. The before-after estimator also requires the absence of aggregate changes in outcomes due to the macroeconomy or other factors. If the economy heats up and all units do better, then the estimator will incorrectly assign the gains to the treatment.

In terms of the notation, let *t* denote a post-program period and *t'* a pre-program period. The before-after estimator is given by

$$\Delta_{BA} = E(Y_{1t} \mid D = 1) = E(Y_{1t} - Y_{0t'} \mid D = 1).$$

The estimator is consistent only if $E(Y_{0t} \mid D = 1) = E(Y_{0t} - Y_{0t'} \mid D = 1)$, which is just the stability condition in the previous paragraph expressed in notation. Note that non-random selection of units into treatment does not pose any problem so long as the stability condition is satisfied, because the before-after estimator compares treated units to their earlier selves. The before-after estimator estimates the mean impact of treatment on the treated, and does not capture any general equilibrium effects unless they are included in the outcomes used to calculate the estimates (as with, *e.g.,* city-level outcome variables). Interrupted time series designs, called event history analyses in the empirical finance literature, generalize the before-after estimator by including additional periods of data before and/or after the treatment of interest. This allows the researcher to control more extensively for pre-existing trends in outcomes.

Concerns about confusing treatment effects with general changes in the economy motivate the so-called "difference-in-differences" estimator. This estimator compares the before-after change of treated units with the before-after change of untreated units. In so doing, any common trends, which will show up in the outcomes of the untreated units as well as the treated units, get differenced out. In terms of our notation, the difference-in-differences estimator consists of

$$\Delta_{DD} = [\, E(Y_{1t} \mid D = 1) = E(Y_{0t'} \mid D = 1)\,] - [\, E(Y_{0t} \mid D = 0) = E(Y_{0t'} \mid D = 0)\,].$$

The common time trend assumption that justifies the estimator is given by:

$$E(Y_{0t} \mid D = 1) = E(Y_{0t'} \mid D = 1) - E(Y_{0t} \mid D = 0) = E(Y_{0t'} \mid D = 0)\,]$$

Researchers most commonly estimate the difference-in-differences model using a relatively simple regression model, as in

$$Y_i = \beta_0 + \beta_T T_t + \beta_D D_i + \beta_{DD} T_i D_i + \varepsilon_i,$$

where $T_i = 1$ in period "t" and $T_i = 0$ otherwise and where we omit the *X* variables (from the notation – not from the model) for simplicity. The coefficient β_T measures the effect of the common time trend, while β_D estimates the time invariant difference in untreated outcomes between the

treated and untreated units, and β_{DD} provides the difference-in-differences impact estimate. In a common effect world (which is a world with homogeneous treatment effects), β_{DD} estimates the mean impact of treatment on the treated (and all of the other mean impact parameters). In a heterogeneous effects world, BDD estimates the mean impact of treatment on the treated under fairly general assumptions. Like the before-after estimator, the difference-in-differences estimator generalizes to the case of more time periods either before or after the treatment or both. Blundell, *et al*. (2001) discuss and implement such estimators.

Panel data models constitute the most general version of these estimators. These models apply to data sets with multiple observations over time on many treated units (and perhaps some untreated units). A regression is run of the outcome variable of interest on exogenous covariates plus dummy variables for each unit and each time period. The unit dummy variables control for permanent differences in outcomes among units, just as in the simple difference-in-differences model. The time period dummies control for aggregate effects in each period. Panel models require some variation in the timing of the treatment; without such variation, the treatment effect cannot be distinguished from the aggregate time effects.[4] These models also require that the timing of treatment among units not depend on transitory changes in outcomes. This is not an innocuous assumption. As Heckman and Smith (1999) show in the context of a government job training program, individual participation depends critically on transitory labor market shocks, with the result that longitudinal estimators have a large bias.

In terms of the notation introduced earlier, the basic panel model has the following form:

$$Y_{it} = \beta_0 + \beta_D D_{it} + \mu_i + \mu_t + \varepsilon_{it},$$

where β_D is the panel data impact estimator, D_{it} is a time-varying indicator for treatment, μ_i is a unit-specific intercept, μ_t is a time-period-specific intercept and I again omit the *X* for simplicity. The time period intercepts soak up any common trends, while the unit-specific intercepts soak up time invariant differences between units. What is left, essentially, is a weighted average of before-after estimates for the different treated units.

A couple of examples illustrate how the models work. Evans and Topoleski (2002) evaluate the impact of Native American casinos on employment and other outcomes (such as crime and bankruptcy in the same county) using panel data methods. They construct panel data on outcomes for Native American tribes in the US that do and do not have casinos. They combine this with data on the timing of casino openings for those tribes that have them. Because of the complicated legal structure surrounding casino gambling in the US, and the fact that both the state and federal governments play a role, there

is a lot of variation in both the incidence and timing of casinos among tribes. Moreover, the variation in the timing of casino openings among tribes that have them is plausibly unrelated to the temporal pattern of outcomes in the absence of the casinos. Put more simply, the timing depends on the vagaries of the political and legal systems, and not on changes in the tribal employment rate. The tribal dummy variables included in the model take account of permanent differences between tribes, and the year dummies take account of nationwide trends affecting all tribes. They find that casinos increase tribal employment to population ratios as well as increasing employment in the surrounding county (as well as crime and bankruptcies).

Coates and Humphreys (1999) provide another economic development application of panel data methods. They compile panel data on cities in the US, including the presence or absence in each year of professional football, baseball and basketball franchises and the construction of new stadiums and arenas. Using the same basic framework described above, they examine the effect of pro sports franchises and the construction of new sports facilities – almost always with substantial public money and almost always justified as engines of economic growth – on both the levels and growth rates of economic activity. They find little in the way of economic effects from public spending on pro sports teams, which comports with the remainder of the serious literature on the topic; see, *e.g.,* the papers collected in Noll and Zimbalist (1997).

Overall, panel data methods represent a powerful tool when longitudinal data are available on treated and untreated units, when the timing of treatment varies among units, and when the timing of treatment is unrelated to the outcomes being studied, conditional on the included conditioning variables. It is this latter condition that sometimes gets ignored in the literature. Additional data allow the testing of this assumption in many contexts, see, *e.g.,* Moffitt (1991) and Heckman and Hotz (1989). For readers interested in the potential pitfalls of panel methods, the highly controversial literature on the impacts of state "right to carry" laws (these allow certain classes of US citizens to carry concealed weapons), which relies almost entirely on such methods, digs deep into what can go right, and what can go wrong, with this approach. See, *e.g.,* Lott and Mustard (1997), Black and Nagin (1998) and Ayres and Donohue (2002).

3.4. Selection on unobservables: instrumental variables

Methods based on instrumental variables (IV) or exclusion restrictions represent an alternative econometric strategy for dealing with selection on unobservables. An instrument, or exclusion restriction, is a variable that affects participation in the treatment but does not affect outcomes other than through its effect on treatment participation (conditional on the other variables included in the outcome equation). The "exclusion restriction" usage

arises from the fact that such variables can be excluded from the outcome equation but included in the treatment equation. Unlike longitudinal methods, IV methods require only cross-sectional data, and they can potentially deal with selection on unobservables that vary over time. See Angrist and Krueger (2001) for a longer non-technical discussion of IV methods and see Angrist and Krueger (1999) and Heckman, LaLonde and Smith (1999) for more technical treatments.

A simple example with a binary instrument in a common effects world provides the basic idea. Consider two otherwise identical towns, one of which is close to a public training center and the other of which is far away from the same training center. Each town includes 100 eligible persons, of whom 50 have a car. The outcome in the absence of training equals 100 for all eligibles. The benefit of training (net of the opportunity cost of participant time) equals 10 for everyone (a common effect world), while the cost for those in the near town to get to the training center equals 1. For the eligibles in the far town, the cost of transport to the training center equals 5 for those with a car and 15 for those without a car. The upshot of all this is that all of the eligibles in the near town take the training but that only the eligibles with cars in the far town take training. As a result, we can use location as an instrument, because it affects participation in the treatment – a different fraction of the eligibles participates in each town – but does not affect outcomes other than through its effect on treatment – everyone gets 100 in the absence of treatment in both towns.

The Wald estimator for binary instruments, given by

$$\Delta_{IV} = \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{Pr(D = 1 \mid Z = 1) - Pr(D = 1 \mid Z = 0)}$$

suffices for our simple example (though two-stages least squares could also be used and yields an equivalent estimate). In the Wald estimator, Z denotes the instrument and D denotes participation in treatment as before; in our simple example, $Z = 1$ for eligibles in the near town and $Z = 0$ for individuals in the far town. Plugging in the numbers from the example yields

$$\Delta_{IV} = \frac{[100 + (1.0)(10)] - [100 + (0.5)(10)]}{1.0 - 0.5} = \frac{5}{0.5} = 10,$$

which is the correct answer. This example illustrates the key point that by inducing variation in treatment receipt unrelated to outcomes in the absence of treatment, the instrument identifies the treatment effect through comparisons of the outcomes of groups with different values of the instrument.

When the instrument is continuous or multi-valued rather than binary, or when multiple instruments are available (which represents good fortune indeed), standard two-stage least squares methods replace the Wald estimator.

In the two-stage estimator, the "endogenous" variable (participation in treatment) is first regressed on the instrument or instruments and any exogenous X variables in the model. In the second stage, the outcome is regressed on the predicted values from the first stage as well as the X Intuitively, using the predictions from the first stage, rather than the participation dummy itself, omits variation in participation not resulting from factors unrelated to outcomes in the absence of treatment. A small literature in econometrics explores less restrictive semi-parametric evaluation estimators based on exclusion restrictions, although these estimators have seen very little use in applied work. See Newey, Power and Walker (1990) and Blundell and Powell (2001) for further discussion.

Instruments are a wonderful tool when available, but where do they come from? Deliberate creation represents the most direct way of obtaining instruments. One way to think about social experiments is that they are devices to create good instruments; this includes the randomized encouragement design described in Section 3.1. See Heckman (1996) for more on this view of experiments. A second form of deliberate instrument creation consists of theory combined with clever data collection. For example, Card (1995) adds data on the distance to the nearest college or university to a standard individual-level data set and then uses distance as an instrument for years of schooling. The final, and in practice the most important, form of data collection combines theory with institutional knowledge. Institutional changes that seem unrelated to the outcomes of interest (or at least whose timing is not related to them, as with, *e.g.,* some court decisions) but affect participation in a treatment can provide good instruments. The bus strike used to provide variation in pre-natal care in Evans and Lien (2002) provides such an example. Differences in program management choices or in program intensity across jurisdictions constitute a potential source of instruments in many economic development contexts. In general, evaluators should think of finding instruments as a side-benefit of thinking about the economics of a given evaluation problem and of learning about the relevant institutions, both necessary activities in their own right.

The bivariate normal selection model of Heckman (1979) is closely related to the instrumental variables model. The bivariate normal model, as its name suggests, assumes that the error terms in the outcome and participation equation have a bivariate normal distribution. Under this assumption, Heckman (1979) provides a two-stage estimator that estimates the impact of treatment on the treated in a common effect world.[5] Formally, the bivariate normal model does not require an exclusion restriction – the functional form assumptions suffice to identify the parameter of interest. However, it is well known in the literature (see, *e.g.,* the survey by Puhani, 2000) that the model tends to instability without an exclusion restriction. As such, evaluators

should avoid this estimator unless they have a good instrument at hand. More generally, as Heckman and Robb (1985) note, this estimator makes stronger assumptions than the IV estimator; if the common effects assumption is plausible in a given context, the IV estimator is preferred.

Things become more complicated in a world with heterogeneous treatment effects, particularly when those treatment effects are correlated with the instrument. To see this, return to the simple example of the two towns sharing one training center considered above. Suppose that half the persons eligible for the program in each town have a benefit of 10 from the program, while half have a benefit of five. Suppose too that transport costs are now homogenous, so that everyone in the near town has a cost of zero to getting to the training center, while everyone in the far town has a cost of seven. In this version of the story, everyone in the near town again participates in training, along with half of the eligibles in the far town. Now, however, instead of the eligibles in the far town with low transport costs participating, it is the eligibles in the far town with the impacts of 10 who participate. For them, the impact of 10 exceeds the transport cost of seven, while for the eligibles with an impact of five in the far town the transport cost of seven makes it not worth their while to participate.

How do these changes in the story of the two towns affect the estimate produced by the Wald estimator? The mean outcome in the near town is now 107.5 = [100 + (0.5 * 10) + (0.5 * 5)]. The mean outcome in the far town equals 105 = 100 + (0.5 * 10). The probability of participation in the near town equals 1.0 and that in the far town equals 0.5. Thus, the Wald estimate equals (107.5 – 105)/(0.5) = 2.5/0.5 = 5.0. This estimate might seem surprising, given that most of those who participate (two thirds to be exact) receive an impact of 10.

The key feature of this version of the story lies in the correlation between the impacts (conditional on participation) and the instrument. In the near town, the mean impact among participants is 7.5, while in the far town it is 10. When a binary instrument such as this one is correlated with the heterogeneous treatment effects conditional on participation, the Wald estimator no longer estimates the mean impact of treatment on the treated. Instead, it estimates what Imbens and Angrist (1994) call a Local Average Treatment Effect (LATE). In simple terms, the Wald estimator now estimates the mean impact on those units who change their participation status in response to the change in the value of the instrument. The units who change their participation status when the instrument changes in the story of the two towns are those with an impact of five; they participate in the near town but not in the far town. Thus, the LATE in this case equals five.

Note that the other standard treatment parameters do not equal five in this case. For example, the mean impact of treatment on the treated equals

(2/3) * 10 + (1/3) * 5 = 25/3 = 8.33. The average treatment effect equals 7.5, because by assumption half of the eligibles in each town have an impact of 10 and half have an impact of five. The relationship between the treatment on the treated and the LATE is instructive here. The treatment on the treated parameter exceeds the LATE because the inframarginal participants – those who participate regardless of the value of the impact (in this case, those with an impact of 10) have higher average impacts than the marginal participants, as economic theory would predict. The three parameters differ in this case and they answer different policy questions of interest. Heckman (1997), Angrist and Krueger (1999) and Heckman, LaLonde and Smith (1999) discuss the binary instrument case in greater detail.

The paper by Sweetman, Warburton, McPhee and Warburton (2003) provides a nice applied example of LATE estimation. In their study, the instrument consists of the person who makes the final decision on disability benefit cases in the Canadian province of British Columbia. When the person making the decision changes, and the acceptance probability increases, the authors can estimate a very interesting LATE – namely the impact of receiving disability payments on the marginal candidates whose applications would get rejected under one regime but get accepted under the other. They cannot, of course, estimate the impact of receiving disability insurance payments on the inframarginal applicants whose cases would be approved under both regimes.

Things get a bit more complicated with continuous instruments, or with multiple instruments, or with continuous (or multi-valued) treatments in a heterogeneous treatment effects world. See Angrist and Imbens (1995) and Heckman and Vytlacil (1998) for discussions of multi-valued treatments. The bivariate normal model also generalizes to the heterogeneous treatment case, and has received a lot of attention in the literature. In this case, the additional structure provided by the strong distributional assumptions on the error terms allows the estimation of numerous LATEs, as well as the ATE and the treatment on the treated parameter. See Heckman, Tobias and Vytlacil (2003) for further details and an empirical example. Recently, a semiparametric version of the bivariate normal selection model has appeared in the literature. A full description of this model is beyond the scope of this paper; see Heckman and Vytlacil (2001b) for details.

In sum, instruments represent a powerful tool for deriving compelling estimates of the impacts of local economic development programs when they are available. To date, IV methods have seen little use in the economic development literature; as such, clever researchers who seek out novel sources of exogenous variation likely have some low-hanging empirical fruit to harvest that would add to both to our knowledge of the impacts of economic development policies and to our knowledge of the performance of these estimators in practice.

3.5. *Discontinuity designs*

Discontinuity designs apply to treatments allocated using a particular variable or set of variables, with the treated and untreated units distinguished by a sharp break in the value of the variables. Thus, we might imagine a program that provides remedial education to pre-school students based on their score on a standardized test, with those below a certain score receiving the treatment and those above it not. Or, grants to small towns might be made available to only those towns with populations below 15 000. In general, suppose that treatment is provided to units with a value of some variable Z greater than a cutoff C.

Absent additional assumptions, discontinuity designs estimate the mean effect of treatment on units located at the cutoff C by comparing the outcomes of units just above the cutoff to the outcomes of units just below the cutoff. In terms of our notation, the discontinuity design estimates

$$\Delta_{DIS} = E(Y_1 - Y_0 \mid X \gg C).$$

For example, in the case of the grants to small towns, the discontinuity estimator estimates the effect of the grants on towns with populations of approximately 15 000 by comparing the outcomes of towns with populations just under 15 000 with the outcomes of towns with populations just over 15 000. The exact form of the comparison depends on the particular discontinuity estimator selected.

Of course, in a common effect world, the impact estimated in the discontinuity design generalizes to other units. Adding the assumption that the untreated outcome is linear in covariates leads to the so-called "regression discontinuity" design, which can be estimated using standard regression methods. Depending on the application, discontinuity designs may capture general equilibrium effects. Whether they do or not depends on the extent to which outcomes in the treated units influence outcomes in the untreated units. For example, in the case of the grants to small towns, the estimator will pick up any general equilibrium effects captured by town-level outcome variables, but will be biased by migration from large to small towns in response to the grants.

The key to the discontinuity design is that units (or others) must not be able to manipulate Z so as to cause certain units to be treated and others not. Some treatments will meet this requirement and others may not. The example given above of a treatment rule based on a standardized test with no subjective component is an example of the former. A policy that provides benefits to firms with fewer than five employees provides an example of the latter. A sufficiently attractive benefit will cause some firms with six or more employees to cut their payroll down to five. These firms will almost certainly be a non-random sample of firms with six or more employees. As a result,

comparisons at the margin between firms with five and six employees no longer estimate the parameter of interest.

The literature provides some useful theoretical discussions. See, *e.g.,* the related section in Heckman, LaLonde and Smith (1999) as well as, at a more technical level, Hahn, Todd and van der Klaauw (2001). The latter paper highlights the tradeoff between bias and variance associated with deciding how much weight to assign to observations at some distance from the cutoff point. Applications include van der Klaauw (2002), who looks at the effect of financial aid on university admissions, Black (1999), who looks at housing values along the borders of school attendance areas (within school districts) and Pence (2003), who looks at the effects of state mortgage regulations along state borders.

3.6. General equilibrium methods

Four types of econometric evaluation methods seek to directly estimate general equilibrium effects. First, as noted in the preceding sections, standard partial equilibrium evaluation methods can capture general equilibrium effects in cases where the unit of analysis incorporates these effects. Thus, for example, if a treatment is randomly assigned to some towns and not to others, and the general equilibrium effects occur within towns rather than between them, than comparing town-level outcomes between towns randomized into the experiment and towns randomized out of the experiment will capture the general equilibrium effects.

The second method consists of traditional (some might say "old style") multiple equation models that link various aspects of local economic development together. These models are similar in spirit to the multiple equation macroeconomic models used by banks and others to generate short-term economic forecasts, but with more specific case study assumptions built in regarding the particular local economic development context. These models have fallen out of favor in the academic literature because they violate many of the rules of sound econometric practice, as the equations often consist of one endogenous variable regressed on several others, with no instruments in sight. The theoretical basis underlying the multiple-equation system typically lacks much in the way of formal structure. The presence of many equations often makes it difficult to see where estimated effects come from and, in practice, the models often require some subjective input to produce reasonable numbers. In short, these models are hard to like, but the demand for numbers, combined with the lack of simple alternatives keep them in play for practical applications.

The third method is what I call the "magic multiplier" method. This method plays a leading role in evaluations of transit projects and sports

infrastructure investments, where the consulting firm performing the evaluation usually has a clear idea of the desired result in advance. In this method, some measure of direct impacts is constructed in a more or less reasonable way. The direct impacts then get multiplied by the magic multiplier, which is supposed to capture all the spillover effects. The particular values chosen for the magic multiplier in a given application typically have little formal justification, which indeed is one of their attractions to those performing the evaluation and their clients. See the papers in Noll and Zimbalist (1997), as well as Crompton (1995), for more details.

The fourth approach to directly estimating general equilibrium effects consists of specifying a structural general equilibrium model of the relevant economy that includes the program, calibrating or estimating the parameters of the model, and then simulating the model with and without the program.[6] Structural general equilibrium models make very strong assumptions about how the different elements of the economy affect one another. Given the complexity of these models, the analyst has no choice but to treat very simply all but those aspects of the economy most relevant to the issue at hand. While the strong assumptions and structure represent a disadvantage in one sense, they represent the strength of these models as well. All the assumptions are clear and on the table, and whatever effects emerge from the model can be traced to particular aspects of the model economy and thereby back to the underlying theory. Because of their complexity and because of the expense associated with these models, undertaking the development of such a model makes sense only for large programs (either in terms of expenditure or in terms of the fraction of the relevant units directly treated), where the answer matters a lot and where we expect important general equilibrium effects.

The literature contains a handful of such evaluations; I highlight four notable examples here, all drawn from the field of active labor market policy. Davidson and Woodbury (1993) construct a general equilibrium search model that allows them to evaluate the general equilibrium effects of the US Unemployment Insurance (UI) bonus programs.[7] These programs, whose partial equilibrium impacts were estimated by a series of social experiments, paid UI claimants a cash bonus if they found work within the first 11 weeks of their benefit claim. Davidson and Woodbury's (1993) model indicated that from 30 to 60 per cent of the partial equilibrium impact gets cancelled out in general equilibrium due to displacement and changes in the optimal search effort of unemployed workers not eligible for the bonus. Lise, Seitz and Smith (2003) modify Davidson and Woodbury's (1993) model to evaluate the Canadian Self-Sufficiency Program, a wage subsidy to long-term income assistance recipients who find full-time work. They find that general equilibrium effects, including changes in optimal search effort and in the distribution of wages, change the social cost-benefit performance of the

program from positive to negative. Blundell, Costa-Dias and Meghir (2003) use general equilibrium methods to examine a wage subsidy program in the UK and find that taking account of general equilibrium effects makes a big difference to their findings. Finally, Heckman, Lochner and Taber (1998, 1999) consider the general equilibrium effects of a $500 per year tuition subsidy to university attendance in the US. They find that taking account of changes in equilibrium skill prices means that the estimated general equilibrium impacts are smaller than the partial equilibrium impacts by a factor of ten. In each of these cases, taking general equilibrium effects into account plays an important role in getting the correct answer about the effects of a policy change.

3.7. Choosing among alternative non-experimental evaluation methods

When feasible, an experimental evaluation will provide the most compelling evidence on the effectiveness of local economic development programs. When an experiment is not feasible, the evaluator must choose among the alternative non-experimental evaluation methods summarized in Sections 3.2 to 3.6.

The lucky analyst enters the process at the beginning, and can influence the program design and implementation as well as the data collection with a specific estimation strategy (or strategies) in mind. In this happy situation, the analyst can build in two or three alternative evaluation strategies, thereby providing multiple lines of evidence and allowing for the (not unlikely) possibility that one of them will not work out in practice. The unlucky analyst enters at the end of the process, and must try to choose an evaluation method that fits data and institutions chosen by others.

The literature provides a lot of guidance to both the lucky and the unlucky analyst, guidance that frequently gets ignored in evaluation practice. Most of this guidance comes from a growing list of papers that use experimental data sets to benchmark the performance (usually in terms of bias) of alternative non-experimental estimators in different contexts defined by the available data and the institutional setup of the program, where the latter in turn determines the nature of the process by which some units come to receive treatment. This literature includes LaLonde (1986), Fraker and Maynard (1987), Heckman and Hotz (1989), Bell, Orr, Blomquist and Cain (1995), Friedlander and Robins (1995), Heckman, Ichimura and Todd (1997), Heckman, Ichimura, Smith and Todd (1998), Heckman and Smith (1999), Dehejia and Wahba (1999,2002), Agodini and Dynarksi (2001), Glazerman, Levy and Myers (2003), Michalopoulos, Bloom and Hill (2004) and Smith and Todd (2004). The literature also includes a simulation study, Section 8.3 of Heckman, LaLonde and Smith (1999) that examines the performance of alternative non-experimental estimators for various data generating processes.

The early parts of this literature framed the question of interest in terms of finding the one true estimator – the magic bullet that would slay the beast of selection bias in every context. More recently, the literature has realized that this is not the right question, because there is no magic bullet. As described above, different non-experimental evaluation strategies make different assumptions about the nature of the selection process and about the available data. When those assumptions hold, a given estimator will provide consistent estimates of certain parameters of interest. When they do not, it will not. Thus, rather than looking for one single estimator that works universally, the literature now seeks the mapping, or relationship, between the institutions and data available in a given context (and the parameter of interest) and the choice of a non-experimental evaluation strategy. Sometimes, as in Hui and Smith (2003), the data available in a given context do not support *any* estimator.

The literature that makes use of experimental benchmarks teaches a number of somewhat obvious (though not so obvious that they have not been ignored in many published papers and even more unpublished ones) but important lessons. A few examples serve to demonstrate the value added by this line of research. The evidence from the series of papers using the National Supported Work Demonstration experimental data show that the handful of variables (age, race, education and lagged annual earnings) available in the US Current Population Survey (CPS) do not suffice to control for selection into a program that served a highly disadvantaged population including ex-convicts and ex-addicts. The "selection on observables" strategies described in Section 3.2 require rich data on observable determinants of participation and outcomes. Heckman, Ichimura, Smith and Todd (1998) demonstrate the importance of drawing comparison group members from the same local labor markets as participants when evaluating active labor market programs. They also show that using outcomes measured in different ways for treated and untreated units (such as administrative data for one and survey data for the other) can lead to outcome differences that look like selection bias but really constitute systematic measurement differences. Heckman and Smith (1999) show that when individuals select into a treatment based on transitory labor market shocks, as they do in most active labor market policies, longitudinal estimator strategies such as those described in Section 3.3, which assume selection based on permanent differences, fare quite poorly. Heckman and Hotz (1989) demonstrate the value of statistical specification tests in contexts with rich enough data to allow their use.

While the specific lessons from the literature derive mainly from active labor market policies, the general lessons hold when choosing a non-experimental estimator for all sorts of economic development programs. A sound evaluation will pay close attention to the nature of the institutions that determine

selection into treatment. These institutions determine the nature of any selection bias, and thereby the plausibility of particular non-experimental evaluation strategies. A sound evaluation will also pay close attention to the fit between the available data and the particular evaluation method employed. Matching methods make no sense without rich data. Instrumental variable methods make no sense without a good instrument. Longitudinal methods make no sense with cross-sectional data or when selection into treatment depends on transitory, rather than permanent, characteristics. In short, a sound evaluation builds on economic theory, econometric theory and existing evidence in choosing a non-experimental evaluation strategy that matches the data and institutions present in a given context.

4. Alternatives to econometric evaluation

4.1. Participant self-evaluation

Evaluators could save a lot of time, money and econometric effort if program participants could reliably evaluate a program directly. In the case of a program that treats individuals, this consists of asking participants, following their participation, whether or not the program made them better off and, if so, how much. A simple survey question replaces all the econometric issues discussed in Section 3. Indeed, many evaluations of US employment and training programs include such questions, and the performance standards for the US Workforce Investment Act include customer satisfaction measures; see US Department of Labor (2000). A similar procedure could apply to firms as well, with the relevant officer chosen to answer the question as in Bartik (2004). Survey methods could even be used to get at some sorts of general equilibrium effects, as is done in the literature on the valuation of environmental amenities; see, *e.g.,* Portney (1994). For example, local residents could be asked how much they value their new small business incubator, even if they do not make use of it themselves.

While participant self-evaluation sounds good in theory, surprisingly little direct evidence exists regarding its ability to get the correct answer. In order for participant self-evaluation to yield valid impact estimates, respondents have to correctly estimate the unobserved counterfactual of what would have happened to them had they not participated, and then compare it to their realized experience as participants.

Heckman and Smith (1998) present some direct evidence on the validity of participant self-evaluation that I reproduce here in Table 12.1. Experimental treatment group members in the US National Job Training Partnership Act Study were asked in a follow-up survey 18 months after random assignment if they thought the program benefited them (see table notes for exact question wording). Table 12.1 shows the fraction of each of the four demographic groups

Table 12.1. **Self-assessments of JTPA impact: experimental treatment group (National JTPA study, 18-month impact sample)**

	Adult males	Adult females	Male youth	Female youth
	61.63 (0.81)	68.10 (0.68)	62.62 (1.29)	66.29 (1.09)
Full-sample percentages Percentage who self-report participating				
Percentage of self-reports participants with positive self-assessments	62.46 (1.04)	65.21 (0.85)	67.16 (1.59)	71.73 (1.29)
Overall percentage with positive self-assessments	38.49 (0.81)	44.41 (0.73)	42.06 (1.32)	47.55 (1.16)
Percentage of self-reported participants with a positive self-assessment by primary treatment received				
None (dropouts)	48.89 (2.07)	51.44 (1.85)	58.90 (3.33)	61.56 (2.79)
Classroom training in occupational skills	74.10 (2.15)	73.47 (1.36)	72.73 (3.60)	75.28 (2.30)
On-the-job training at private firm	75.13 (2.18)	78.90 (2.14)	71.00 (4.56)	75.00 (4.04)
Job-search assistance	59.57 (2.27)	59.80 (2.18)	68.09 (3.94)	68.94 (4.04)
Basic education	62.96 (4.67)	56.55 (3.84)	70.97 (4.09)	78.44 (3.19)
Work experience	66.67 (9.83)	68.75 (5.84)	82.76 (7.14)	73.17 (7.01)
Others	58.47 (3.65)	66.40 (2.98)	62.50 (4.77)	77.98 (3.99)

Notes:
1. Reported proportions are based on responses to the question "Do you think that the training or other assistance you got from the program helped you get a job or perform better on the job?" This question was asked only of self-reported participants within the treatment group. The overall fraction of positive self-assessments assumes that self-reported non-participants would have provided negative self-assessments.
2. The primary treatment is the one in which the trainee participated for the most hours according to the administrative records of the JTPA sites. Most trainees received only one service; few received more than two. See Smith (in press) for a detailed discussion. Note that for some self-reported participants the JTPA administrative records indicate that no services were received.
3. (3) Estimated standard errors in parentheses.

Source: Table 8.11 of Heckman and Smith (1998).

in the experiment – adult males, adult females, male youth and female youth – who answered yes. It compares these to the experimental impact estimates on self-reported earnings over the 18-month period between random assignment and the follow-up interview. Table 12.1 reveals little correlation between the fractions self-reporting that they benefited from the program and the experimental earnings impact estimates, which suggests that program participants do not do a good job of constructing the counterfactual necessary to determine whether or not the program made them better off. See Heckman and Smith (1998) or Smith and Whalley (2004) for further discussion of the evidence from the JTPA experiment.

The indirect evidence also suggests trouble for participant self-evaluation. First, the discussion in Section 3 makes it clear that researchers have great difficulty constructing reasonable estimates of average counterfactuals – probably an easier task than estimating the counterfactual for a single person. The literature on behavioral decision theory suggests that individuals have all sorts of cognitive problems with less difficult tasks such as making consistent decisions when choices are presented in different ways, and that most people are poor "intuitive statisticians." Survey effects may also cause trouble. Respondents may not want to risk offending an interviewer by saying that a program did not help them (or they may not want to admit to themselves that they wasted their time and energy on it!). See, *e.g.,* Bradburn, Sudman and Wansink (2004) in regard to interviewer effects of this sort. Finally, as Bartik notes in his chapter in this volume, for programs involving monetary transfers (or valuable in-kind transfers), respondents can have a direct interest in the program continuing, and so may report a behavioral response even when one does not exist in order to keep the goodies coming.

While the limited available evidence argues against relying on participant self-evaluation for impact analysis, this does not preclude gathering useful information from participants about other aspects of their participation (or, indeed, from gathering useful information from eligible non-participants about why they chose not to take part). For example, participants will likely have a good sense of the quality of service they received and of the amount of red tape involved in participation.

4.2. Performance standards

Administrative performance standards represent another potentially inexpensive alternative to impact analysis. Performance standards typically consist of quantitative measures of program outputs (the number of checks sent out on time) or outcomes (how many of the trainees found a job within a month after finishing the training program). In terms of our notation, they generally consist of functions of Y_1. They have grown in popularity as part of the "reinventing government" movement of the 1990s – see, *e.g.,* Osborne and Gaebler (1992) – and now pervade the US government as a result of the Government Performance and Results Act (GPRA) of 1993.

Performance standards have many uses, and in some contexts they tell you all you need to know. For example, the primary mission of the US Social Security Administration (SSA) consists of sending out checks (or making direct deposits) to the correct people at the correct time and in the correct amount. A performance measure that gives the fraction of the time that this happens tells much (if not all) of what needs to be told about how well SSA performs. Their task consists of an outcome, rather than an impact, and so is well suited to management using outcome based performance measures. Of course, the

checks SSA sends out will have behavioral impacts of interest to economists and policy-makers – such impacts are not well captured by performance measures based on outcome levels.[8]

In contexts where program impacts represent the main object of concern, reliance on performance standards as a proxy for impact estimates requires evidence of a systematic relationship between the two. To make the problem concrete, consider a job training program that uses "entered employment rates" – such as the fraction of trainees employed 90 days after leaving the program, as its performance measure. This measure is a version of Y_1. In order for the performance measure to provide a useful proxy for the program's impact – that is, for the difference it makes in the employment rate relative to what would have occurred had the participants not participated in the program – the performance measure must be positively correlated with the program impact. In this case, this means that employment after leaving the program must be correlated with the change in employment status induced by the program. In one extreme case, that where no one finds employment without the program, the correlation equals one and the performance measure equals the impact. More generally, there is no particular reason why this condition should hold.

A program that is most effective for those clients least likely to find employment on their own may exhibit a negative relationship between employment rates and impacts. To see this, consider the extreme case of easy-to-serve clients who would always find employment on their own and hard-to-serve clients who never would, but do so half the time when they participate in the program. In one month, the program serves half of each type of client. Its employment rate is 0.75 because the easy to serve all find employment and so do half of the hard to serve, while its impact is 0.25, reflecting the fact that it only benefits hard to serve clients. In another month, it serves only hard-to-serve clients. In that month, its employment rate is 0.50 but its impact is also 0.50, because none of the hard-to-serve clients would have found employment on their own. For this program, the employment rate performance measure is negatively related to program impacts, and so provides a poor proxy for program impacts (and a strong incentive for program managers to cream skim by serving only easy-to-serve clients).

The evidence on this question for US employment and training programs, such as that presented in Heckman, Heinrich and Smith (2001) and Barnow (1999) and summarized in Barnow and Smith (2003) suggests that common performance measures used in that context, such as entered employment rates, do not correlate very well, if at all, with program impacts.

Very similar issues arise in other program contexts when performance gets judged according to outcomes rather than impacts. For example, government programs that subsidize commercial research and development are often

judged based on the fraction of projects that pay off, which provides an incentive for program operators to choose projects that would have been funded anyway without the subsidy, rather than funding projects with (potentially) large spillovers but low private benefits. It is the latter type of project that provides the economic justification (as opposed to the political justification) for the subsidy. See Wallsten (2000a) for a popular discussion of the US Small Business Innovation Research (SBIR) program and Wallsten (2000b) for a more academic discussion. Wallsten's research suggests that each dollar of the funds provided under the SBIR program crowds out a dollar of private research funds. This results from a focus on choosing projects likely to succeed in the market, as this is the metric used to judge program success. His findings indicate that the program has no impact on the total amount of research undertaken. Overall, the literature suggests that performance standards do not represent a general substitute for econometric impact evaluation.

5. Practice

This section briefly considers some important practical issues associated with evaluating local economic development policies. The first subsection focuses on when not to do an evaluation. The second subsection focuses on how to choose an evaluator and then how to evaluate an evaluation once completed, and the third highlights some important issues in cost-benefit analysis. This section builds on the discussion in Smith and Sweetman (2001).

5.1. *When not to do an evaluation*

Evaluations consume time and resources. As such, evaluations, like the programs being evaluated, should go forward only when their benefits are likely to exceed their costs. This subsection outlines a number of situations in which an evaluation will likely not pass a standard cost-benefit test and where, as a result, the money that would be spent on evaluation would be better spent on other things.

The first situation where an evaluation is a bad idea is when money is short and other basic administrative functions are not in order. Before doing an evaluation, program operators should have a clear idea of which units participate in their program, whether or not the participating units are eligible for the program and, in a voluntary program, how the participating units compare to the population of all eligible units. They should also know how much money the program is spending, what it is being spent on, and which treated units the money is being spent on. Collecting and examining all this information represents a basic fiduciary duty on the part of the program operator acting as an agent to the longsuffering taxpayer (or to those who donate to a non-profit organisation that sponsors economic development

projects). These fiduciary duties should come before attempts at evaluation; after all, a program that is not under control in terms of eligibility and costs seems unlikely to produce much in the way of impacts.

The second situation where evaluation can be skipped is when the impact is known in advance. This can happen for two reasons. First, there may already be a substantial, high quality evaluation literature for a particular type of program. For example, thanks to a long series of experimental evaluations at the state level, the literature has a pretty clear idea of the effects of both mandatory and voluntary job search assistance programs on single mothers on welfare in the United States. See, *e.g.,* Gueron and Pauly (1991) and Bloom and Michalopoulos (2001). Additional evaluations, unless they cover a non-trivially different treatment modality or population, probably do not justify their cost. The second way for the answer to be known in advance is when programs really exist just to transfer money to some favored individuals, firms or other interests, with the economic development justification serving as a useful distraction for a bored public and an inept media. Subsidies to particular large firms seeking to locate a new plant represent an important example of such programs. From a national point of view, such bidding wars between states and localities can do no better than have a zero impact, and will have a negative one to the extent that they cause geographic misallocations of production. Indeed, the European Union forbids competition of this type among its member nations.[9] These programs also undermine the rule of law, an important determinant of long-run growth at the international level, by treating some firms differently just because of their size, mobility, or political connections. A bit of cynicism, combined with simple economic theory and careful attention to where the money goes, usually suffices to identify such circumstances. Sadly, exposing them to the light of day does not always (or even often) result in their disappearance.

The third situation where an evaluation may not represent a good investment arises when the samples available for the evaluation would lack the size required for statistical inference. For example, a program that subsidizes only five firms will not provide a clear statistical picture of program impacts using any of the methods outlined in Sections 3.1 to 3.6. The econometric methods outlined here require a substantial number of treated and untreated units in order to identify impacts statistically distinguishable from zero. Limited sample sizes can also arise from budgetary limitations in situations with a large number of units potentially available for an evaluation, but with substantial costs (*e.g.,* survey costs) of adding each unit to the data. Statisticians have developed formal methods for determining the number of observations required to detect an impact of a given size with a certain degree of confidence under various assumptions about the variance of the outcome measure and other characteristics of the evaluation context. Applying these

methods constitutes a "power analysis"; such an analysis should precede the decision about whether to proceed with an evaluation except in cases with very small or very large sample sizes. See, *e.g.,* Maxwell (2000) for a discussion of power analysis and additional citations.

The fourth situation that calls for not doing an evaluation arises when the data do not exist to support an evaluation, or when high quality data cannot be obtained at a cost within the available budget. For example, many relatively expensive evaluations of major government programs often rely on survey data with surprisingly (or even shockingly) low response rates. In such situations, the evaluation must devote additional attention to issues of selective non-response, which cast doubt on the consistency and generality of the findings. In such situations, it might be better to spend enough money to get a reasonable response rate (say, 80 per cent) or to use an alternative data source, such as administrative data. The latter are, of course, no panacea. Data quality has a low priority at many agencies, which means an evaluation must devote substantial resources to *ex post* cleaning of the data, perhaps discarding some fields entirely. See, *e.g.,* Hotz and Scholz (2002) for further discussion of administrative data. In general, there exists some reservation data quality level below which an evaluation becomes valueless.

A lack of evaluation expertise constitutes the final situation when no evaluation may dominate some evaluation. If the organisation potentially undertaking the evaluation lacks access to the funds or the personnel to carry it forward, or to evaluate it when done, then it should generally not attempt the evaluation, particularly if the literature contains relatively strong evaluations of similar programs operated in similar contexts. Weak evaluations do not justify the money they cost, which leads directly to the topic of the next subsection, which concerns how to choose an evaluator and how to evaluate the evaluator's evaluation.

5.2. *Choosing and evaluating an evaluator*

Anyone can declare herself an evaluator and seek contracts for performing evaluations. In practice, many individuals (and collections of individuals in firms or project-specific coalitions) perform evaluations, including economists, statisticians, psychologists and sociologists. Some do evaluation on the side, in addition to teaching and/or academic research; others do evaluation full time. Some know all the latest econometric methods while others can only run regressions. The particular evaluation context and the budget loom large here, so I offer only a few general observations.

First, experiments are harder than you think; if you want to do one and have not done one before, hire a firm that knows how to do it. Such firms include MDRC, Mathematica and Abt Associates, to name a few. Second,

different types of evaluators have different characteristics, which should be matched to the needs at hand. Professional evaluation firms cost the most, but they have a lot of experience and deliver a very polished product on time and, generally, within budget. Academics, on the other hand, tend to cost a bit less, and sometimes know more econometrics, but have a lower probability of finishing on time and a bit less polish.

Third, some evaluators will take your money and give you back an embarrassing mess. For example, the paper by Gregory (2000), published (for unknown reasons) in the journal *Evaluation*, suggests an evaluation centered on a variant of the "sites of oppression matrix". In this approach to evaluation, key stakeholders sit around and contemplate all the ways in which life has treated them poorly (or at least those ways somehow related to the program) and the evaluator then writes about the filled-in matrix. Rather obviously, such exercises represent an entertaining diversion for the stakeholders and easy money for the evaluator, but yield no insight about the impact of the program. Fourth, sometimes you can get the econometric part of an evaluation done at low cost if you provide an academic researcher with interesting data that they can use to write articles for publications in scholarly journals. Indeed, many if not most published evaluations of social programs were not paid for by the agencies operating the program they evaluate.

Evaluations, like programs, require evaluation. Some evaluations are very good while others are very weak. Not all agencies that commission evaluations have the internal staff expertise to undertake such evaluations. Even if they do, external quality checks may add substantial value to the evaluation and also increase its credibility. A number of methods exist for incorporating external feedback and review into the evaluation process. Large-scale evaluations often include a technical review panel of experts who provide feedback at critical stages, such as the design report and the draft impact analysis. In smaller evaluations, a single outside expert may play this role by providing comments on drafts of various reports. Once an evaluation is complete, feedback is still useful to guide readers in determining how much weight to place on the results obtained. Encouraging publication in peer-reviewed journals is one way to accomplish this; and the knowledge that the final product will eventually be sent out for peer review provides an incentive for quality throughout the evaluation process. Inclusion of reviewer comments as an appendix to the published final report, as in Jacobson and Petta (2000) plays a similar role.

5.3. Cost-benefit analysis

Evidence-based policy builds on a foundation of serious and thorough cost-benefit analyses of various policy alternatives. Cost-benefit analyses, in turn, rest on a foundation of high-quality econometric program evaluations. Many

trees have given their lives for books on cost-benefit evaluation. This subsection does not attempt a general treatment, but instead highlights a few key issues that have received too little attention in the literature. The discussion here draws on the discussion in Section 10.2 of Heckman, LaLonde and Smith (1999).

Impact evaluations commonly generate net impact estimates for a short time, usually a few months or years. Yet, at least in certain contexts, such as human capital development programs or infrastructure investments, we expect impacts to persist for some time. In such contexts, two related issues arise. The first issue is how to project the estimated benefits outside the period of the data. While theory or evidence from other evaluations of similar programs with longer follow-up periods can play a role in guiding this decision, in the end the best course will likely consist of constructing estimates of the cost-benefit performance of the program assuming multiple plausible durations for program benefits. The recent experimental Job Corps evaluation does not do this, and as a result, particularly in its executive summary provides a somewhat misleading guide to policy. That evaluation presumes as a base case that program impacts last (essentially) forever and concludes on that basis that the program easily passes a cost-benefit test (see Table 3 in Burghardt *et al.* (2001) and the surrounding discussion). Yet because of the high cost of this program, and the relatively short period of post-program data collection, without the assumed future benefits the program would have a positive gross impact (which represents a major achievement relative to most government employment and training programs for youth) but would fail the cost benefit test miserably.

The second issue is what discount rate to use for future benefits. A small literature exists that attempts to estimate optimal social discount rates under various assumptions [see, *e.g.,* the discussion and references in Liu (2003)]. Once again, reporting cost-benefit estimates for multiple plausible rates seems best.

Another issue in cost-benefit analysis concerns the deadweight cost of taxation, called the "excess burden" in the public finance literature. This number measures the cost to the economy of the marginal tax dollar including (ideally) both the direct costs of operating the tax system and the indirect costs of the distortions induced by the tax system. The literature offers a wide variety of estimates of this cost, ranging from only a few cents to well over one dollar per dollar of tax revenue. See, *e.g.,* Browning (1987) and Snow and Warren (1996) for further discussion and evidence. Once again, given the uncertainty in the literature, presenting multiple estimates based on different values seems the best course (and ignoring the deadweight costs altogether, as too many evaluations do, seems the worst course).

In the case of each of the cost-benefit issues considered here, presenting multiple estimates that rely on different assumptions does two important things. First, it allows readers with different prior beliefs than the evaluator

about these issues to see the cost-benefit estimates under his or her preferred assumptions. Second, it highlights to policymakers the range of cost-benefit estimates consistent with the data. This forces them to confront uncertainty about program performance and at the same time provides a sense of the robustness of any recommendations concerning the policy.

6. Conclusion

The resources devoted to local economic development programs have valuable alternative uses. Econometric program evaluations play a key role in determining when to continue with economic development programs and when to shut them down. In this chapter, I have emphasized five key themes in evaluating such programs.

First, and perhaps foremost, I have emphasized the importance of reading the literature. We have learned a lot about how to do econometric policy evaluations in the last two decades, but this knowledge has not yet affected evaluation practice to the extent that it should. The knowledge we have gained includes both advances in methods, as well as advances in practice, including the use of administrative data and clever identification strategies.

Second, there is no magic bullet. No econometric evaluation estimator provides consistent estimates for all (or even most) possible combinations of data, institutions and parameter of interest. Regression does not do this, matching does not do this, the bivariate normal model does not do this, difference-in-differences does not do this, and IV does not do this. The search for such an estimator, which animated the literature for many years, has now come to an end, replaced by a more sensible research program designed to identify the mapping between characteristics of the data and institutions and the parameter of interest to the estimators likely to yield consistent answers.

Third, heterogeneous treatment effects matter. They affect the choice and interpretation of econometric evaluation estimators. They imply careful thought about the exact parameter of interest required to answer a particular policy question. The conceptual literature is advancing rapidly in this area, but has already revolutionized how evaluators think about what they do.

Fourth, general equilibrium effects matter in many evaluation contexts, particularly when considering local economic development programs, which generally aim to create such effects. The potential presence of general equilibrium effects has important implications for the joint decision regarding which econometric evaluation estimator to employ and the unit of analysis for the evaluation. Depending on the unit of analysis, some estimators will miss or, worse still, be biased by, general equilibrium effects. Pick the unit of analysis too large, and program effects get lost in the shuffle; pick the unit of analysis too small and general equilibrium effects get missed. General equilibrium effects

seem to attract bad evaluation practices as well, particularly the use of magic multipliers in evaluations of public infrastructure investments.

Finally, not every program will benefit from an evaluation. Before proceeding with one, some thought, some power calculations, and an informal cost-benefit analysis of the evaluation itself will help to sort out situations where an evaluation represents a sound investment from situations where it represents a waste of time and money.

Notes

1. Of course, adding some technical skills would not only represent personal development on the part of notation-averse economic developers, it might also improve their ability to promote local economic development.
2. For example, Stata reports marginal effects equal to derivatives evaluated at the mean of the covariates, as opposed to the slightly more difficult, but technically preferable, procedure of calculating mean derivatives by taking the average of derivatives evaluated at the covariates values for each observation.
3. In the technical literature, this is called the "curse of dimensionality". See Smith and Todd (2004) for further discussion.
4. The literature that uses panel models to evaluate the impact of the switch from Aid to Families with Dependent Children to Temporary Aid to Needy Families in the US provides a good example of the dangers of using these models with very little variation in the timing of treatment. This literature relies on limited monthly variation in the implementation of TANF across states – variation that seems likely to be related to the outcomes under study and therefore violates the assumptions that justify panel models.
5. The model can also be estimated in one stage using full information maximum likelihood methods. While there is an efficiency gain from doing so, the two-stage version may be more robust to misspecification, as it relies less strongly on the joint normality assumption.
6. The debate regarding the relative merits of calibration and estimation of structural equilibrium models lies well beyond the scope of this paper. See Hansen and Heckman (1996), Kyland and Prescott (1996) and Sims (1996) for three relatively non-technical presentations of different views of the debate.
7. See Meyer (1995) for an overview of UI-related policy experiments.
8. I have taken this example from Wilson (2000), a book well worth reading for those interested in why managing and evaluating government programs proves so difficult in practice.
9. See the discussion and references in Bartik (2003b).

References

AGODINI, Roberto and Mark DYNARSKI (2001), "Are Experiments the Only Option? A Look at Dropout Prevention Programs". Mathematica Policy Research Working Paper No. 8723-300.

ANGRIST, Joshua and Guido IMBENS (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity". *Journal of the American Statistical Association*. 90(430): 431-432.

ANGRIST, Joshua and Alan KRUEGER (1999), "Empirical Strategies in Labor Economics". In Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics, Volume 3A*. Amsterdam: North-Holland. 1277-1366.

ANGRIST, Joshua and Alan KRUEGER (2001), "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments". *Journal of Economic Perspectives*. 15(4): 69-85.

AYRES, Ian and John DONOHUE (2002), "Shooting Down the More Guns, Less Crime Hypothesis". NBER Working Paper No. 9336.

BARNOW, Burt (1999), "Exploring the Relationship between Performance Management and Program Impact: A Case Study of the Job Training Partnership Act". *Journal of Policy Analysis and Management*. 19(1): 118-141.

BARNOW, Burt and Jeffrey SMITH (2003), "Performance Management of US Job Training Programs". Unpublished manuscript, University of Maryland.

BARTIK, Timothy (2003a), "Local Economic Development Policies". W.E. Upjohn Institute Staff Working Paper No. 03-91.

BARTIK, Timothy (2003b), "Thoughts on American Manufacturing Decline and Revitalization". *Employment Research*. 10(4): 1-4.

BARTIK, Timothy (2004), "Evaluating the Impacts of Local Economic Development Policies On Local Economic Outcomes: What Has Been Done and What is Doable?" In this volume.

BELL, Stephen, Larry ORR, John BLOMQUIST and Glen CAIN (1995), *Program Applicants as a Comparison Group* in *Evaluating Training Programs*. Kalamazoo: W.E. Upjohn Institute for Employment Research.

BLACK, Dan and Daniel NAGIN (1998), "Do Right-to-Carry Laws Deter Violent Crime?" *Journal of Legal Studies*. 27(1): 209-219.

BLACK, Dan and Jeffrey SMITH (2004), "How Robust is the Evidence on the Effects of College Quality? Evidence from Matching". *Journal of Econometrics*, forthcoming.

BLACK, Dan, Jeffrey SMITH, Mark BERGER and Brett NOEL (2003), "Is the Threat of Reemployment Services More Effective than the Services Themselves? Evidence from Random Assignment in the UI System". *American Economic Review*. 93(4): 1313-1327.

BLACK, Sandra (1999), "Do Better Schools Matter? Parental Valuation of Elementary Education". *Quarterly Journal of Economics*. 114(2): 577-599.

BLOOM, Dan and Charles MICHALOPOULOS (2001), *How Welfare and Work Policies Affect Employment and Income: A Synthesis of Research*. New York: Manpower Demonstration Research Corporation.

BLOOM, Howard (1984), "Accounting for No-Shows in Experimental Evaluation Designs". *Evaluation Review*. 82(2): 225-246.

BLUNDELL, Richard and Monica COSTA DIAS (2000), "Evaluation Methods for Non-Experimental Data". *Fiscal Studies*. 21(4): 427-468.

BLUNDELL, Richard and Monica COSTA DIAS (2002), "Alternative Approaches to Evaluation in Empirical Microeconomics". *Portuguese Economic Journal*. 1(1): 91-115.

BLUNDELL, Richard, Monica COSTA DIAS and Costas MEGHIR (2003), "The Impact of Wage Subsidies: A General Equilibrium Approach". Unpublished manuscript, University College, London.

BLUNDELL, Richard, Monica COSTA DIAS, Costas MEGHIR and John VAN REENAN (2001), "Evaluating the Impact of a Mandatory Job Search Assistance Program". IFS Working Paper No. WP01/20.

BLUNDELL, Richard and James POWEL (2001), "Endogeneity in Semiparametric Binary Response Models". CEMMAP Working Paper No. CWP05/01.

BRADBURN, Norman, Seymour SUDMAN and Brian WANSINK (2004), *Asking Questions: The Definitive Guide to Questionnaire Design*. Jossey-Bass.

BROWNING, Edgar (1987), "On the Marginal Welfare Cost of Taxation". *American Economic Review*. 77(1): 11-23.

BURGHARDT, John, Peter SCHOCHET, Sheena MCCONNELL, Terry JOHNSON, Mark GRITZ, Steven GLAZERMAN, John HOMRIGHAUSEN and Russell JACKSON (2001), *Does the Job Corps Work? Summary of the National Job Corps Study*. Princeton, NJ: Mathematica Policy Research.

CALMFORS, Lars (1994), "Active Labor Market Policy and Unemployment – A Framework for the Analysis of Crucial Design Features". *OECD Economic Studies*. 22(1): 7-47.

CARD, David (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling". In Louis Christofides, Kenneth Grant and Robert Swidinsky, eds., *Aspects of Labour Market Behavior: Essays in Honor of John Vanderkamp*. Toronto: University of Toronto Press. 201-222.

COATES, Dennis and Brad HUMPHREYS (1999), "The Growth Effects of Sport Franchises, Stadia and Arenas". *Journal of Policy Analysis and Management*. 18(4): 601-624.

CROMPTON, John (1995), "Analysis of Sports Facilities and Events: Eleven Sources of Misapplication". *Journal of Sports* Management. 9(1): 14-35.

DAVIDSON, Carl and Stephen WOODBURY (1993), "The Displacement Effects of Reemployment Bonus Programs". *Journal of Labor Economics*. 11(4): 575-605.

DEHEJIA, Rajeev and Sadek WAHBA (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs". *Journal of the American Statistical Association*. 94(448): 1053-1062.

DEHEJIA, Rajeev and Sadek WAHBA (2002), "Propensity Score Matching Methods for Non-Experimental Causal Studies". *Review of Economics and Statistics*. 84(1): 151-161.

DOOLITTLE, Frederick and Linda TRAEGER (1990), *Implementing the National JTPA Study*. New York: Manpower Demonstration Research Corporation.

EICHLER, Martin and Michael LECHNER (2002), "An Evaluation of Public Employment Programmes in the East German State of Sachsen-Anhalt". *Labour Economics*. 9(2): 143-186.

EVANS, William and Diana LIEN (2002), "The Benefits of Prenatal Care: Evidence from the PAT Bus Strike". Unpublished manuscript, University of Maryland.

EVANS, William and Julie TOPOLESKI (2002), "The Social and Economic Impact of Native American Casinos". NBER Working Paper No. 9198.

EBERTS, Randall and Christopher O'LEARY (2004), "Evaluating Training Programs: Impacts at the Local Level". In this volume.

FRAKER, Thomas and Rebecca MAYNARD (1987), "The Adequacy of Comparison Group Designs for Evaluation of Employment-Related Programs". *Journal of Human Resources*. 22(2): 194-227.

FRIEDLANDER, Daniel and Philip ROBINS (1995), "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods". *American Economic Review*. 85(4): 923-937.

FRÖLICH, Markus (2004), "Finite Sample Properties of Propensity Score Matching and Weighting Estimators". *Review of Economics and Statistics*. Forthcoming.

GLAZERMAN, Steven, Dan LEVY and David MYERS (2003), "Nonexperimental Versus Experimental Estimates of Earnings Impacts". *Annals of the American Academy of Political and Social Science*. 589: 63-93.

GREENBERG, David and Mark SHRODER (1997), *Digest of Social Experiments, 2nd Edition*. Lanham, Maryland: Rowman and Littlefield.

GREENE, William (2002), *Econometric Analysis, 5th Edition*. Upper Saddle River, NJ: Prentice Hall.

GREGORY, Amanda (2000), "Problematizing Participation: A Critical Review of Approaches to Participation in Evaluation Theory". *Evaluation*. 6(2): 179-199.

GUERON, Judith and Edward PAULY (1991), *From Welfare to Work*. New York: Russell Sage.

HAHN, Jinyong (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects". *Econometrica*. 66(2): 315-331.

HAHN, Jinyong, Petra TODD and Wilbert VAN DER KLAAUW (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design". *Econometrica*. 69(1): 201-209.

HANSEN, Lars and James HECKMAN (1996), "The Empirical Foundations of Calibration". *Journal of Economic Perspectives*. 10(1): 87-104.

HECKMAN, James (1979), "Sample Selection Bias as a Specification Error". *Econometrica*. 47(1): 153-161.

HECKMAN, James (1996), "Randomization as an Instrumental Variable". Review of Economics and Statistics. 78(2): 336-341.

HECKMAN, James, Carolyn HEINRICH and Jeffrey SMITH (2001), "The Performance of Performance Standards". *Journal of Human Resources*. 37(4). 778-811.

HECKMAN, James, Neil HOHMANN, Jeffrey SMITH and Michael KHOO (2000), "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment". *Quarterly Journal of Economics*. 115(2): 651-694.

HECKMAN, James and V. Joseph HOTZ (1989), "Choosing Among Alternative Methods of Estimating the Impact of Social Programs: The Case of Manpower Training". *Journal of the American Statistical Association*. 84(408): 862-874.

HECKMAN, James, Hidehiko ICHIMURA, Jeffrey SMITH and Petra TODD (1998), "Characterizing Selection Bias Using Experimental Data". *Econometrica*. 66(5): 1017-1098.

HECKMAN, James, Hidehiko ICHIMURA and Petra TODD (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme". *Review of Economic Studies*. 64(4): 605-654.

HECKMAN, James, Hidehiko ICHIMURA and Petra TODD (1998), "Matching as an Econometric Evaluation Estimator". *Review of Economic Studies*. 65(2): 261-294.

HECKMAN, James, Robert LALONDE and Jeffrey SMITH (1999), "The Economics and Econometrics of Active Labor Market Programs". In Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics,* Volume 3A. Amsterdam: North Holland. 1865-2097.

HECKMAN, James, Lance LOCHNER and Christopher TABER (1998), "General Equilibrium Treatment Effects: A Study of Tuition Policy". *American Economic Review*. 88(2): 281-386.

HECKMAN, James, Lance LOCHNER and Christopher TABER (1999), "General Equilibrium Cost-Benefit Analysis of Education and Tax Policies". In Gustav Ranis and Kakshmi Raut, eds., *Taxes, Growth and Development: Essays in Honor of Professor T.N. Srinivasan*. New York: Elsevier. 291-349.

HECKMAN, James and Salvador NAVARRO-LOZANO (2004), "Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models". *Review of Economics and Statistics*. Forthcoming.

HECKMAN, James and Richard ROBB (1985), "Alternative Methods for Evaluating the Impact of Interventions". In James Heckman and Burton Singer, eds., *Longitudinal Analysis of Labor Market Data*. New York: Cambridge University Press for Econometric Society Monograph Series. 156-246.

HECKMAN, James and Jeffrey SMITH (1995), "Assessing the Case for Social Experiments". *Journal of Economic Perspectives*. 9(2): 85-110.

HECKMAN, James and Jeffrey SMITH (1998), "Evaluating the Welfare State". In Steiner Strom, ed., *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial*. Cambridge University Press for Econometric Society Monograph Series. 241-318.

HECKMAN, James and Jeffrey SMITH (1999), "The Pre-Programme Dip and the Determinants of Participation in a Social Programme: Implications for Simple Programme Evaluation Strategies". *Economic Journal*. 109(457): 313-348.

HECKMAN, James, Jeffrey SMITH and Nancy CLEMENTS (1997), "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Program Impacts". *Review of Economic Studies*. 64(4): 487-537.

HECKMAN, James, Jeffrey SMITH and Christopher TABER (1998), "Accounting for Dropouts in Evaluations of Social Programs". *Review of Economics and Statistics*. 80(1): 1-14.

HECKMAN, James, Justin TOBIAS and Edward VYTLACIL (2003), "Simple Estimators for Treatment Parameters in a Latent Variable Framework". *Review of Economics and Statistics*. 85(3): 748-754.

HECKMAN, James and Edward VYTLACIL (1998), "Instrumental Variable Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling when the Return Is Correlated with Schooling". *Journal of Human Resources*. 33(4): 974-987.

HECKMAN, James and Edward VYTLACIL (2001a), "Identifying the Role of Cognitive Ability in Explaining the Level of and Change in the Return to Schooling". *Review of Economics and Statistics*. 83(1): 1-12.

HECKMAN, James and Edward VYTLACIL (2001b), "Local Instrumental Variables". In Cheng Hsiao, Kimio Morimune and James Powell, eds, *Nonlinear Statistical Modeling: Essays in Honor of Takeshi Amemiya*. Cambridge: Cambridge University Press.

HECKMAN, James and Edward VYTLACIL (2004), "The Econometric Evaluation of Social Programs". In James Heckman and Edward Leamer, eds. *Handbook of Econometrics, Volume 6*. Amsterdam: North-Holland. Forthcoming.

HOLLISTER, Robinson, Peter KEMPER and Rebecca MAYNARD (1984), *The National Supported Work Demonstration Project*. Madison: University of Wisconsin Press.

HOTZ, V. Joseph and Karl SCHOLZ (2002), "Measuring Employment and Income Outcomes for Low-Income Populations with Administrative and Survey Data". In *Studies of Welfare Populations: Data Collection and Research Issues*. National Research Council: National Academy Press. 275-315.

HUI, Shek-Wai and SMITH, J. (2002), "The Labor Market Impacts of Adult Education and Training in Canada". Report prepared for Human Resources Development Canada.

IMBENS, Guido (2000), "The Role of the Propensity Score in Estimating Dose-Response Functions". *Biometrika*. 87(3): 706-710.

IMBENS, Guido (2004), "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review". *Review of Economics and Statistics*, forthcoming.

IMBENS, Guido and Joshua ANGRIST (1994), "Identification and Estimation of Local Average Treatment Effects". *Econometrica*. 62(4): 467-476.

JACOBSON, Lou and Ian PETTA (2000), "Measuring the Effects of Public Labor Exchange (PLX) Referrals and Placements in Washington and Oregon". Workforce Security Occasional Paper No. 2000-06, Employment and Training Administration, US Department of Labor.

KYDLAND, Finn and Edward PRESCOTT (1996), "The Computational Experiment: An Econometric Tool". *Journal of Economic Perspectives*. 10(1): 69-85.

LALONDE, Robert (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". *American Economic Review*. 76(4): 604-620.

LECHNER, Michael (1999), "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification". *Journal of Business and Economic Statistics*. 17(1): 74-90.

LECHNER, Michael (2000), "An Evaluation of Public-Sector-Sponsored Continuous Vocational Training Programs in East Germany". *Journal of Human Resources*. 35(2): 347-375.

LECHNER, Michael (2001), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption". In Michael Lechner and Friedhelm Pfeiffer, eds, *Econometric Evaluation of Labor Market Policies*. Heidelberg: Physica. 43-58.

LIU, Liqun (2003), "A Marginal Cost of Funds Approach to Multi-Period Public Project Evaluation: Implications for the Social Discount Rate". *Journal of* Public *Economics*. 87(7-8): 1707-1718.

LISE, Jeremy, Shannon SEITZ and Jeffrey SMITH (2003), "Equilibrium Policy Experiments and the Evaluation of Social Programs". Unpublished manuscript, University of Maryland.

LONG, Sharon and Douglas WISSOKER (1995), "Welfare Reform at Three Years: The Case of Washington's Family Independence Program". *Journal of Human Resources*. 30(4): 766-790.

LOTT, John and David MUSTARD (1997), "Crime, Deterrence and Right-to-Carry Concealed Handguns". *Journal of Legal Studies*. 26(1): 1-68.

MAXWELL, Scott (2000), "Sample Size and Multiple Regression Analysis". *Psychological Methods*. 5(4): 434-458.

MEYER, Bruce (1995), "Lessons from the US Unemployment Insurance Experiments". *Journal of Economic Literature*. 33(1): 91-131.

MICHALOPOULOS, Charles, Howard BLOOM and Carolyn HILL (2004), "Can Propensity Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" *Review of Economics and Statistics*, forthcoming.

MOFFITT, Robert (1991), "Program Evaluation with Nonexperimental Data". *Evaluation Review*. 15(3): 291-314.

MOFFITT, Robert (2003), "Remarks on the Analysis of Causal Relationships in Population Research". Unpublished manuscript, Johns Hopkins University.

NEWEY, Whitney, James POWELL and James WALKER (1990), "Semiparametric Estimation of Selection Models: Some Empirical Results". *American Economic Review*. 80(2): 324-328.

NEWHOUSE, Joseph (1994), *Free for All: Lessons from the Rand Health Insurance Experiment*. Cambridge, MA: Harvard University Press.

NOLL, Roger and Andrew ZIMBALIST (1997), *The Economic Impact of Sports Teams and Facilities*. Washington, DC: Brookings Institution.

ORR, Larry (1998), *Social Experiments: Evaluating Public Programs with Experimental Methods*. PLACE: Sage Publications.

OSBORNE, David and Ted GAEBLER (1992), *Reinventing Government: How The Entrepreneurial Spirit is Transforming the Public Sector*. Boulder, CO: Perseus.

PENCE, Karen (2003), "Foreclosing on Opportunity: State Laws and Mortgage Credit". FEDS Working Paper 2003-16.

PORTNEY, Paul (1994), "The Contingent Valuation Debate: Why Economists Should Care". *Journal of Economic Perspectives*. 8(4): 3-17.

PUHANI, Patrick (2000). "The Heckman Correction for Sample Selection and Its Critique". *Journal of Economic Surveys*. 14(1): 53-68.

RAVALLION, Martin (2001), "The Mystery of the Vanishing Benefits: An Introduction to Evaluation". *World Bank Economic Review*. 15(1): 115-140.

ROSENBAUM, Paul and Donald RUBIN (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika*. 70(1): 41-55.

ROSENBAUM, Paul and Donald RUBIN (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score". *American Statistician*. 39: 33-38.

SIMS, Christopher (1996), "Macroeconomics and Methodology". *Journal of Economic Perspectives*. 10(1): 105-120.

SMITH, Jeffrey (2000), "A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies". *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 136(3): 247-268.

SMITH, Jeffrey and Arthur SWEETMAN (2001), "Improving the Evaluation of Employment and Training Programs in Canada". Unpublished manuscript, University of Maryland.

SMITH, Jeffrey and Petra TODD (2004), "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?", *Journal of Econometrics*, forthcoming.

SMITH, Jeffrey and Alexander WHALLEY (2004), "Are Participants Good Evaluators? Evidence from the Job Training Partnership Act". Unpublished manuscript, University of Maryland.

SNOW, Arthur and Ronald WARREN (1996), "The Marginal Welfare Cost of Public Funds: Theory and Estimates". *Journal of Public Economics*. 61(2): 289-305.

SWEETMAN, Arthur, William WARBURTON, Rob MCPHEE and Rebecca WARBURTON (2003), "Disability Status: Impacts on Health and Welfare Dependence". Unpublished manuscript, Queen's University.

US DEPARTMENT OF LABOR (2000), "Core and Customer Satisfaction Performance Measures for the Workforce Investment System". Training and Employment Guidance Letter No. 7-99. Washington, DC: Employment and Training Administration.

VAN DER KLAAUW, Wilbert (2002), "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Design". *International Economic Review*. 43(4): 1249-1287.

WALLSTEN, Scott (2000a), "The R&D Boondoggle". *Regulation*. 23(4): 12-17.

WALLSTEN, Scott (2000b), "The Effects of Government-Industry R&D Programs on Private R&D: The case of the Small Business Innovation Research Program". *Rand Journal of Economics*. 31(1). 82-100.

WILSON, James (2000), *Bureaucracy: What Government Agencies Do and Why They Do It*. New York: Basic Books.

WINSHIP, Christopher and Stephen MORGAN (1999), "The Estimation of Causal Effects from Observational Data". *Annual Review of Sociology*. 25: 659-706.

WOOLDRIDGE, Jeffrey (2001), *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

WOOLDRIDGE, Jeffrey (2002), *Introductory Econometrics: A Modern Approach, 2nd Edition*". Mason, Ohio: South-Western College Publishing.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Asses What Works among Programmes and Policies
© OECD 2004

Chapter 13

Evaluation and Third-sector Programmes

by
Andrea Westall,
Deputy Director New Economics Foundation,
London, United Kingdom

Introduction and policy implications

The third-sector is a crucial element of strategies for local economic development. Not only do the activities of third-sector organisations often contribute to a range of positive regeneration impacts, but they are also increasingly seen to be important players in economic development partnerships and in the delivery of local public services. third-sector organisations argue for a particular set of "added value" impacts such as social capital and community involvement which can be vital elements of economic development. They are also credited with creating innovations which can inform new policies and models of economic development. The intermediate labour market model, for example, was a third-sector response to long-term unemployment.

Assessing the impacts and outcomes of third-sector activity is therefore necessary to find out just how such organisations contribute to regeneration; the extent of their impacts; and how their particular "value added" can be further supported and developed. The results of good evaluation can be used to support ongoing development and improvement in the sector, aid replicability, refine or change government policy, and support appropriate and effective funding and financing of their activities.

However, good evaluation of the third-sector in local economic development currently appears quite limited. This is due to a variety of factors arising from internal capacity constraints; the requirements of different funders which may conflict, be limited or create overload; and the lack of appropriate processes and measurement systems which are able to capture the particular outcomes, often intangible, which different kinds of third-sector organisation may create. Evaluation is also difficult because third-sector organisations often have a range of goals and activities. Their impacts are also hard to untangle from the local context of which they are a part and from the actions of public, private and other third-sector players.

This paper summarises a range of approaches to third-sector evaluation and particularly stresses the need to develop measurement systems which start from the point of view of the objectives of the organisation itself and its needs rather than just the requirements of outside funders and policy-makers. This not only enables more comprehensive evaluations but also provides information which can directly inform further organisational and policy development.

These approaches cover both process (formative) and *ex post* evidence of impact (summative) evaluation, arguing that underpinning both is a need for systematic data collection and new methods of capturing outcomes that are hard to quantify. It is also important to recognise the importance of incorporating the community and users in the evaluation process in order to create improved and relevant measures of outcomes, identify unintended consequences, and better understand cause and effect. Participative evaluation can also create more robust and effective programmes since it contributes to the regeneration process itself by creating buy-in from the local community as well as involving beneficiaries in finding their own ways to achieve outcomes and targets.

There are a variety of implications for policy makers:

- A greater understanding of how third-sector organisations contribute to local economic development can lead to changes in policy, better partnership arrangements and targeted support for organisations with strong impact on key development objectives.
- There is a need to combine methodologies in order to understand "how something works" as well as "whether it works" and "how it can be improved".
- There is a need to focus on outcomes and not outputs in funding schemes for third-sector organisations.
- third-sector organisations require resources and capacity to be able to evaluate their activities appropriately. This may involve allowing funding streams to incorporate finance for appropriate evaluation or the provision of advice on the most effective methods through support agencies.
- Evaluations that are able to capture social and environmental impacts and relate these to public expenditure and public targets could support the creation of new and innovative forms of funding based on social investment or on social payment or incentives schemes.
- If evaluations are able to capture high social returns on investments but low economic returns, this helps create arguments for the need for subsidy if those outcomes are valuable contributions to local development and policy goals.
- The need to recognise that third-sector organisations face a range of evaluation requirements and that it is more beneficial to support the creation of internal data collection systems which can be used to underpin a range of evaluations both for internal and external use.
- Understanding third-sector impacts may enable a "market making" role for the public sector where support can be given to the development of organisations which are able to address multiple policy priorities in a joined-up way.

- Good evaluation can help the public sector reconsider its procurement strategies and contracts in order to capture the multiple outcomes of third-sector providers, for example, by delivering a required service whilst also reducing unemployment. This means that government contractors should be able to deliver and to design contracts which incorporate the kinds of added-value which the third-sector may be able to provide. They may also wish to allocate more resources to supporting the third-sector in its ability to create and access appropriate tenders and for increasing its capacity to deliver.[1]
- A greater understanding of third-sector organisations and the community-based strategies can give more weight to recognizing and capturing the value of approaches to economic development which are less reliant on top-down programmes. Bottom-up and locally appropriate models which engage the community and relevant contributions from the third-sector, private and public sector need to be appropriately evaluated and assessed in order to determine their relative impacts.

The last point raises an important issue which is not tackled in this paper. It is critical to recognise the multiplicity of actors and outcomes in development partnerships and in disadvantaged areas and to measure the synergy between them and their relationship to overall outcomes. More work needs to be done to develop a systems approach to evaluation which can assess the overall impacts of a range of different activities and organisations, but also identify unique contributions from individual players and the interactions between them.

Understanding the third-sector

The third-sector is a contested concept. It may, for example, be equated with the idea of the "social economy", seen by the European Union as comprising CMAF – co-operatives, mutuals, associations and foundations. Others see this classification as limited, ignoring the many examples of not-for-profits or social enterprises that do not fall into easily defined categories.

In effect, the third-sector describes a space which includes a range of more recognised terms such as not-for-profits, voluntary sector, mutuals, social enterprises or community enterprises. These organisations tend to have in common a focus on social or environmental objectives rather than profit-seeking. However, some do create a profit but reinvest this in their activities or distribute to relevant stakeholders rather than external shareholders or the owner-manager.

In reality, this third-sector space is very diverse and has no clearly defined boundaries. It overlaps with the business sector and with the public sector. There are, for example, some very interesting examples of mission-

driven shareholder businesses and organisations that are part community and part public-owned which break down easy distinctions and often provide new and innovative ways of delivering public services.[2]

The third-sector can range from organisations that are fully-grant dependent to those that are fully self-financing and from those that compete in mainstream markets to those that provide non-marketable goods and services. The range of social and environmental goals is extremely diverse as are their governance structures and the different groups of stakeholders who are engaged with each organisation.

This diversity and overlap with other sectors therefore means that it is difficult to create evaluative frameworks that are conceptually different from those used for other purposes. However, they may require some modification to deal with the distinctiveness of third-sector approaches and outcomes.

How does the third-sector support local economic development?

The importance of the third-sector arises from its role in contributing to regeneration in ways which go beyond a reliance on physical renewal and inward investment. It is widely recognised that regeneration and local economic development is a multi-faceted process involving issues of social capital, health and welfare and local culture as well as creating jobs and new businesses. Indeed, regeneration only benefits residents if ways can be found of ensuring that jobs are created for local people and that any wealth creation does not just "leak" out of the area.

The third-sector can have a role in, for example:

- Increasing employment or employability.
- Supporting the creation of new enterprises, for example, through managed workspaces or organisations supporting the creation of new businesses.
- Providing new inclusive ways of doing business, for example, co-operative models.
- Addressing undermet or unmet needs for goods and services, for example, in social housing, food, and finance.
- Addressing social issues such as poor health or the needs of refugees.
- Building local infrastructure – for example, arts, sport facilities, transport solutions.
- Creating social capital and increased community engagement.
- Recycling and developing responses to other environmental needs.

Why measure impact?

The process of measuring outcomes and impacts is a way of showing whether or not organisations achieve the goals that they set themselves or the requirements of funding arrangements. It also enables organisations to understand "how" they are achieving their aims which can support future organisational development, funding applications and aid in replication of models or in contributing to policy change. Innovations in the third-sector often showcase new ways in which social, environmental or economic policy may be affected. Evaluation can also unearth the unexpected outcomes that can arise from particular programmes of work or interventions.

Another reason for evaluation is to create accountability to, and engagement of, key stakeholders. This is necessary not only for legitimacy but also to create a sense of shared ownership and enable beneficiaries and other knowledgeable stakeholders to contribute expertise towards understanding "how" something is working and how it might be improved in particular local circumstances. This means that participative approaches to evaluation can contribute to the effectiveness of the economic development process itself by creating new ways of addressing local needs (often involving the local people themselves) and responding to specific local opportunities for economic and social change. Changes may seem small and simple but can have large impacts. For example, an evaluation of crime levels in an area, both actual and perceived, found that local people believed that the trees adjacent to their houses in a park helped burglars scale their walls undetected. By removing those trees, feelings of safety were increased. (It was not recorded if actual crime levels went down).[3]

Evaluation is also a necessary part of organisational change and management, supporting staff and directors in the development and revision of programmes and in designing future strategies and more effective ways of working.

We are therefore talking here about evaluation as both a formative process approach, with implications for the ongoing development of a programme, as well as its use in a more summative way to demonstrate final impacts. There is no reason, however, why these two approaches should not be combined. Often, data collected for ongoing process evaluation is the same as that required for summative evaluations.

There are, though, certain third-sector managers who are somewhat dismissive of evaluation. For some, it is a luxury that would be a nice thing to do, were it not that scarce resources need to be targeted on delivery and often on day-to-day survival.

It has also been argued that performance measurement may cause organisations to strive for continuous improvement when this may not be appropriate to the third-sector and current practice should be "good enough".[4]

Similarities and differences of third-sector evaluation methods

Measuring the impact of the third-sector is in some ways no different from evaluating any other local economic development programme or activity. It is primarily about the most appropriate tools and techniques. However, evaluation of the third-sector is often required, or driven, by the need to show particular impacts – to show that a certain way of doing things adds value and to explain why this is the case. It is also important to evaluate all the impacts of an organisation, not just specific activities. There may well be synergies between the projects conducted by an organisation, or the way the organisation works may itself create added-value. One example could be that of tenant-owned housing which, in addition to its services, has been shown to create other benefits including increased self-confidence in tenants and involvement in civic activities.[5] Through considering the overall mission of the organisation and its effectiveness, rather than the impacts of individual projects, whole-organisation evaluation can also show gaps in provision or assess the general capacity and effectiveness of staff and processes. Standard programme evaluation frameworks may not be able to capture all the processes at work, or the specific intangibles created by the organisation.

The kinds of specific outcomes we are talking about have been suggested to include:

- Closeness to the community.
- Active engagement of users and beneficiaries.
- Access to and understanding of disadvantaged groups.
- Trust by users in the quality of a service. This can be particularly important for services where it is difficult for the users or relatives of users to judge quality, for example, in elderly care or childcare.
- Providing nonmarketable goods and services.

These different attributes or activities are often assumed or anecdotal rather than shown. They are also not necessarily confined to the third-sector. It is therefore important that they are demonstrated and that organisations have the ability to measure their impacts and improve their performance on the basis of the information that they gather.

Issues in impact measurement and the third-sector

We need first to distinguish between evaluating the impact of a third-sector organisation in particular programmes and evaluating the impact of the

organisation as a whole. There is also the challenge of disentangling the inputs of different organisations within multi-actor partnerships.

We also need to distinguish between processes that the organisation uses, such as engaging stakeholders, and particular outcomes. This is necessary because it could be argued that engaging stakeholders is time consuming, using resources that may better be focused on direct delivery. However, the processes may improve the delivery or be a valuable end in themselves, creating for example, increased social capital, greater community engagement, buy-in to the programme, or increased self-esteem.

Unfortunately, third-sector organisations are not always good at showing their impacts or their particular added value. One of the main reasons for this is the many evaluation requirements arising from programme funders or other finance providers which can include different measures and different timescales for delivery. The result can be management overload, high costs and partial evaluations of an organisation's activities. Funder evaluation requirements are in reality fairly simply based on outputs rather than outcomes and more qualitative analysis.

There is a tendency, therefore, for many third-sector organisations to see funder monitoring and evaluation as more of a burden than a useful process. They feel that external evaluations generally have little use as learning tools and that they do not really communicate the essence of what organisations are doing and how they make that difference.

Other barriers to quality evaluation include:

- Constraints of time, resources, skills and knowledge.
- The complexity of many evaluative approaches.
- The plethora of different available models which means that organisations are unsure as to the quality of different evaluation techniques, or how to go about choosing the most appropriate approaches for their needs. Additionally, multiple evaluation approaches can make comparability and bench-marking difficult between similar organisations.
- Unavailability of appropriate techniques, particularly those which can evaluate intangible outcomes and impacts. Social firms for example are organisations that operate in mainstream markets but employ predominantly disabled people or those with a mental health problem. They are currently looking to develop a measure of their ability to increase positive health outcomes.
- The multiple nature of objectives and outcomes which can require a whole range of different evaluative approaches.
- The different evaluation requirements of different audiences from funders to beneficiaries to management.

- Funders feeling concerned that some of their money is being used for evaluation and not for service delivery.
- Evaluation being seen as a test rather than a tool that can be used to increase effectiveness.

In addition, there are also general problems which arise in measuring regeneration impacts including:

- The length of time for different outcomes or impacts to become apparent following interventions or initiatives.
- The multiplicity of inter-related factors which influence success or failure within a particular context or with different groups of people.
- The difficulty of disentangling impact from broader causal influences in the local area or in the broader macro-environment.

Types of evaluation techniques used

The kinds of evaluative techniques used by third-sector organisations range from qualitative to quantitative and from external or "extractive" evaluation to evaluation which is participatory and may involve users in defining their own evaluative measures. The more participatory the approach, the more likely it is that the measurement itself becomes an integral part of the regeneration process – whether helping to define or refine the programme and changing the nature of involvement or commitment by stakeholder participants.

Participative evaluation – Participatory Monitoring and Evaluation (PMandE) began in the 1970s in the international development field as a result of the recognition that external evaluation with preset criteria was unable to capture how change happens or to create a process which is inclusive and works with the views and aspirations of those most directly affected.*

* More information on Participatory Monitoring and Evaluation can be obtained, for example, in a Policy Briefing from the Institute for Development Studies Website: Policy Briefing, Issue 12, November 1998. *www.ids.ac.uk*.

This approach involves local people in choosing and designing appropriate indicators to measures local outcomes and also in the data collection process. It can therefore contribute to the renewal process through identifying locally important criteria as well as involving local people in identifying solutions. PMandE challenges some of the established ideas of rigorous data collection and analysis. In some cases data collected may be more reliable. For instance, in Merthyr Tydfil in the UK, information collected

by children on crime in a study in 1996 was better than that collected by the police since people were more likely to tell children the truth.[6]

The process can also be made externally valid through choice of relevant indicators and appropriate sampling.

Many third-sector organisations make use of certain evaluation techniques but often in a fairly basic and *ad hoc* way. There is a whole range of evaluative approaches that could be used but these are not necessarily easy to access or tailored to specific needs.

Strengths

- Feeds back into the regeneration process.
- Engages stakeholders and can lead to a greater sense of ownership of and engagement in the process which can lead to increased trust, awareness and community capacity.
- Builds on local knowledge and experience of what works and how it does so which increases knowledge of causality and therefore learning.
- Can lead to better quality information.
- Engages people who are often left out of formal surveys.
- Stimulates action and civic engagement.

Weaknesses

- Can be time and resource intensive.
- Problems related to comparability and robustness.
- Must balance the need for particular indicators which are comparable across areas in order to assess relative impacts with local choice of indicators that are specific to the particular context. The latter can often be invaluable in understanding how something has happened and how it could be improved. An example of a comparable question would be: "Do you feel safe in your area?" and a more specific question would be "Do you feel safe walking through [named] park?" The first question taps general levels of community feelings of safety and the latter a more specific space where interventions might be made, if appropriate or possible.
- Participation is not a given and requires creativity.
- Participation may be biased through involvement of pressure groups and more "powerful" or vociferous residents. More broadly, subjective assessments can be problematic since they can change according to the timing of the evaluation or the context in which it is made.

There are several overviews of evaluation in the third-sector, for example, the Independent Sector's Measure Project which is a survey of outcome measurement techniques used in the US.[7] There is also a Canadian review of evaluation resources by Bozzo and Hall.[8]

There has been a significant amount of work focused on the third-sector and regeneration. A range of approaches and evaluation frameworks have been developed. These have focused on measuring the particular added value of third-sector organisations around issues such as changes in community involvement or social capital.

Kendall and Knapp (1999) set out an approach to performance measurement which includes some of the possible added value of the voluntary sector – choice, participation, advocacy and innovation – alongside criteria often used to evaluate the public sector – economy, effectiveness, efficiency and equity.[9] Whilst theoretical, this work was used by Community Evaluation Northern Ireland to develop a framework for measuring social capital which could be used to select appropriate indicators.[10]

Another example of a third-sector evaluation framework is that developed by the Scottish Community Development Centre – ABCD – Achieving Better Community Development.[11] This model sees evaluation as an integral part of community development, and community involvement as a vital part of evaluation. They developed a pyramid of outcomes from personal empowerment and participation through to the creation of sustainable communities.

The New Economics Foundation has also developed a framework – Prove It! – for measuring social capital which allows the use of both locally defined measures as well as others which can be compared across areas and projects. The information can therefore contribute to both process and summative evaluations.

Prove It!

Prove It! is an approach to evaluation developed in partnership with Groundwork – an environmental not-for-profit organisation in the UK – and Barclays Bank in order to assess the change in "social capital" in an area as a result of regeneration projects. Whilst social capital is a contentious concept, this project was specifically designed to find out whether projects to improve the local physical environment were also creating change in the community through, for example, increasing the number of relationships between people, or levels of trust. The model is easily accessible to local people as well as the public sector and uses the language of measurement and learning rather than evaluation, which can be discouraging to some.

Prove It! measures the often invisible effects of regeneration activities on local people who are actively involved in choosing indicators and collecting the data. Some indicators are readily available and others are newly designed to capture changes in attitudes and action, for example, in levels of graffiti. Qualitative and quantitative techniques are used, including surveys and existing local data. The full process and its outputs are set out in an easy-to-use handbook.[12]

Unintended impacts of the projects assessed included greater community engagement, ideas for new projects, commitment to the regeneration activities, and enjoyment and learning from the process itself. There was also increased understanding between older and younger generations as a result of the latter being involved in data collection.

Prove It! is being further developed to make the approach more rigorous and to permit benchmarking across activities and areas whilst retaining the participative methodology.

Using indicators

A core element of all the approaches mentioned above is that of evaluating processes and outcomes which are often hard to measure. Examples of such outcomes include increased trust and decreased fear of crime. In order to prevent reinvention of the wheel and to aid comparison, robust and tested indicators are useful. Some of these have already been developed, whilst others are being designed or still need to be created.

A good example of an innovative and useful indicator is that of the measure of "distance travelled" used by many intermediate labour market organisations and employment projects.[13] This concept refers to the progress that a beneficiary makes towards employability, or outcomes that are more easily measured, as a result of the project intervention. It acts as a measure of the soft outcomes achieved such as interpersonal skills, organisational skills, time management, or confidence rather than just looking at whether or not someone has found a job. For many people who have been in long-term unemployment, obtaining a job will be an immense step and considerable effort and time will be needed for them to become employable. Increased "employability" therefore involves a range of skills and attitude changes and, for some people, the distance to be travelled towards achieving employment will be greater than for others.

Such a measure is therefore more appropriate as an indicator of the success of an intervention than just whether or not someone obtains a job. It is also vital to ensuring that funders recognise the real added-value and do not create easy output targets which could bias behaviour so that organisations are pushed towards only dealing with people who are the easiest to employ.

Core outcomes measured relate to key work skills such as communication skills; attitudinal skills such as increased levels of motivation or confidence; personal skills, such as improved timekeeping, and practical skills such as in filling forms. There may then be target-group-specific outcomes for different people, for example, those with learning disabilities.

A study by the then Department for Education and Employment (DfEE) in the UK found that whilst the approaches to measuring "distance travelled" are useful, there are many different techniques in use. This makes comparability difficult.[14] There are also difficulties with the subjectivity of assessors and beneficiaries, as well as problems attributing these changes solely to a project. The result of the DfEE review was to prepare clearer guidance on measurement, and to recognise the need for targeted approaches. This example clearly illustrates the possibilities and inevitable challenges of creating indicators.

The New Economics Foundation has summarised a range of potential indicators, and their uses, both in a publication on community indicators and in the *Prove It!* manual noted above. There is also a useful website which sets out 450 indicators that have been tried and tested worldwide.[15] The New Economics Foundation has recently developed a specific indicator, the Local Multiplier 3, to enable businesses, third-sector organisations and government to see the impact of their economic activities and their contribution to the local economy.

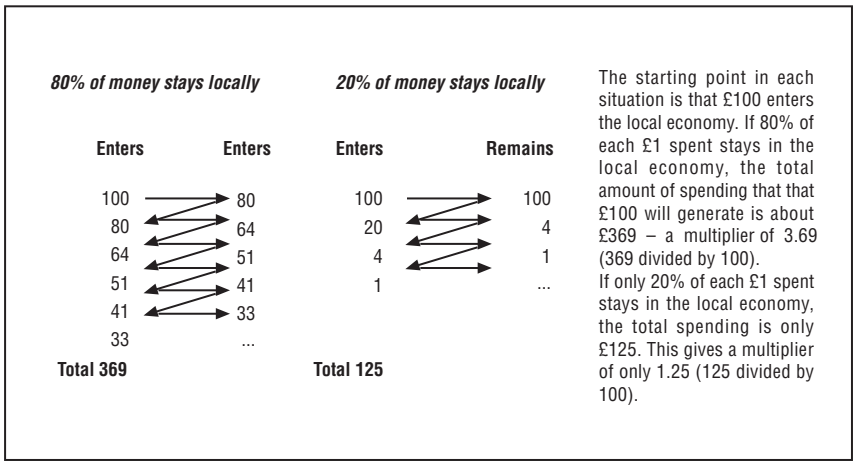
Local Multiplier 3 – LM3

The multiplier can be illustrated through the following example (see Figure 13.1). Let's assume that £100 enters the local economy. If 80% of each £1 stays in the local economy, the total amount of spending that that £100 will generate, over five rounds of spending, is £369. In this case the multiplier will be 3.69. This is found by dividing the total amount of spending by the initial input. If however only 25% of each £1 stays in the local economy the total spending after five rounds is only £125. This will give a multiplier of 1.25 (125 divided by 100).

For simplicity, the LM3 only uses three rounds of spending. However, more can be used, and extrapolation of spending rates can generate more exact measures.

In practice, you first analyse the income of the organisation OR activity and track where it goes using business accounts or questionnaires to find out how money is spent locally and how much is spent outside the defined local area. You then go to the next level of spending and see how the money received by employees or suppliers is spent. Round 1 is the income; Round 2 is the spending; and Round 3 is the spending by staff and suppliers. You then

Figure 13.1. **An example of the local multiplier**



calculate the LM3 by adding up all three rounds and dividing by the original income.

We can illustrate this process through two third-sector examples:

Eden Community Outdoors is a social enterprise which provides opportunities for young people, for example through setting up new businesses or through environmental projects. They calculated their local multiplier to be 2.02, *i.e.* for every £10 received, the company gains an additional £10.20.

Ibstock Community Enterprises is a multi-activity regeneration organisation in a rural community. They wished to measure the impact of a cash-machine which they had installed to respond to the closure of the local bank branch. They found that local people spent between 50-67% of their money within the local area. This point illustrates the importance of looking for interventions which help to bring money into an economy and help it to be spent on local goods and services.

This tool is of particular interest to third-sector organisations because it enables them both to show the economic impact of their activities and also to find practical ways to increase their "multiplier". The organisation can find ways in which the money spent by the organisation can be further channelled

into local activities and therefore contribute to further job and wealth creation. This is particularly important in disadvantaged areas where it is often extremely difficult to determine whether and how regeneration money is "sticking" in the local community or is leaking out in wages or payments to out-of-area suppliers.

The LM3 is not exact. But it is simple enough for organisations and communities to use to evaluate their impacts. It can also be used to look at the benefits of inward investment or the impacts of local procurement decisions by local government or by business. Another example would be social or income benefits payments. When the approach was piloted with Newham Council in London, they found a multiplier of 1.77. In other words, for every £1 of extra benefit take-up, £1.77 more was created in the local community. The implication here is that strategies to increase welfare benefit take-up would help increase wealth and job creation in disadvantaged areas.

More broadly LM3 can be used to show how policies and practices, particularly for regeneration and local economic development, can be changed in order to promote greater local circulation of money or of funding.

The technique is currently being further developed. Its implications are being discussed by policymakers and practitioners in the UK. It is a potentially useful tool for disadvantaged localities or in rural areas where maintenance of local services can often be critically important.

Organisational frameworks for evaluation

There is also a range of frameworks for addressing the multiple nature of organisational outcomes and stakeholder engagement which go beyond outcome measures to addressing organisational processes. All these approaches, however, still require a range of indicators and measures such as those outlined above. They, in effect, create frameworks into which a range of information can be collected and analysed in a variety of ways.

Two approaches are set out below. One is the social audit and the other is an information management system developed by the Roberts Enterprise Development Fund in the US – OASIS.

Social accounting and social audit

Social accounting and social audit began in the early 1970s and was further developed by Traidcraft and the New Economics Foundation in the 1990s to become a voluntary approach rooted in engagement with stakeholders. There is now an international body ISEA – Institute of Social and Ethical Accountability – which promotes best practice and standards which have been used by large corporates such as Shell. The approach is also promoted in several EU-funded transnational programmes.

Social accounting is a framework methodology into which different kinds of impact assessment can fit. The process, put simply, is that you first clearly identify the social objectives and values of the organisation against which any activities are to be assessed. You then define your stakeholder groups and establish social, economic and environmental performance indicators. Some of these indicators are designed by the stakeholders themselves. The accounting process involves collecting subjective and objective data which is then brought together into a social audit, the results verified by an independent assessor, and then published.

Further development of social accounting has enabled smaller third-sector organisations to make use of the model. An approach with wide applicability has been developed for community organisations in Scotland and also the Social Enterprise Partnership has developed the first European Social Audit programme using a Social Audit Toolkit which has been translated into a number of languages.[16]

Social accounting mixes qualitative and quantitative approaches and can increase accountability, empower stakeholders, enable strategic planning and show impacts. It creates a process which can fully engage all relevant groups in identifying needs and solutions.

One major problem with this approach is that it can be quite resource intensive. There is a need for simpler off-the-shelf products. The New Economics Foundation is currently developing, in partnership with several other organisations, a new social audit tool called Ethical Explorer, which will be a simple-to-use online system with appropriate support. The online system provides simple proformas, options for indicators, and analytical tools to reduce the administration and time involved in undertaking a social audit and producing a social report. A prototype is available for view and comment.[17]

There has also been a lot of work in the US, particularly by the Roberts Enterprise Development Fund, to create usable frameworks for evaluating third-sector impact. Their specific focus has been on employment creation models, but the techniques have wider applicability.

OASIS – Ongoing Assessment of Social Impacts

The Roberts Enterprise Development Fund has pioneered a social venture capital approach to working with not-for-profit organisations. They have been developing the capacity of such organisations to measure, track and use social outcome information to more effectively assess the value of their programmes. They would like to further develop their approach to become a new standard for documenting success in not-for-profits.

OASIS enables organisations to track units of service to individuals and also the impact of those services on people over time.[18] Not-for-profits can

assess exactly what services and resources are used together with their impacts, which then enables them to calculate their social return on investment (see the section on *Presenting the Data* below).

In effect, this approach is a customised social management information system which allows organisations to improve their effectiveness and also collect data systematically, which can be used by a variety of funders and evaluators.

The processes involved are not simple. OASIS is highly IT intensive and involves significant restructuring of the organisation itself.

Rubicon

Rubicon is an organisation in California which supports people who are disadvantaged, providing affordable housing, employment, job training, mental health and other support services. Its programmes include support services as well as several social enterprises which employ low income and disabled adults, including a bakery and a homecare business.

Rubicon was spending a great deal of time and money preparing reports and had over 80 funders. However, none of this evaluation activity enabled its funders, or itself, to see the bigger picture. It also had a problem that many of its services were integrated and there was a need for a flexible case management system which could track people and cut down information collection. The development of REDF's CICERO system – Consumer Information Collection, Entry and Reporting for Organisations – took 3 years. However, there were positive impacts. The system led to internal changes, the ability to assess the suitability of programmes to meet needs, improved services and the integration of services. All its contracts and grant reports are now based on this data and it is easier to recognise key indicators of programme success.

Presenting the data

The outcomes of evaluation can be presented in a variety of ways and to a variety of audiences. The social audit is a comprehensive way of summarizing information and presenting it in an accessible format for stakeholders. Elements of the information collected may feed into the development of a variety of benchmarking activities or longer-term rigorous experimental designs.

One interesting approach which is currently being developed in the US and the UK is that of Social Return on Investment.

Social return on investment

This approach enables an organisation to demonstrate both its economic and social return and create a "blended" return which can be used to compare projects, or investments for their relative impacts.

The technique was developed in the US by the Roberts Enterprise Development Fund. In effect, social return on investment (SROI) finds ways to create a monetary value for the social impacts created. An example is the social benefits arising from projects which create employment for disabled people. The social benefits considered here are not only the jobs (and the resultant increased standard of living, self-confidence, independence etc.) but also the improvement in health outcomes as a result of employment. Economic return for an organisation is usually measured as return on investment. Social return is the value of the social, environmental and economic outcomes created by that organisation. Where possible, a proxy for these impacts can be found through estimating future changes in public spending, such as the extent to which tax revenue is increased, or benefit payments decreased.

More specifically, SROI aims to create a measure of the net present value of the stream of future costs and benefits and compare this to the initial investment. The precise way in which this is calculated is set out in a paper by REDF[19]. An example of the SROI in practice is that of Pedal Revolution in San Francisco, which provides young people with training in bike sales and repairs. The SROI process requires creating a measure of the economic and social returns created, and an index which compares the amount of an investment to the value created. An Index of one means that for every dollar (EURO, etc.) invested one dollar (EURO, etc.) of value has been created. If the Index is greater than one then value has been created in excess of the investment. In the Case of Pedal Revolution, the Index of economic return is 9.6. If you include all the extra value created through its social activities the Blended Index of return is 32.5, illustrating the huge relative value created as a result of the social impacts of the organisation.

REDF acknowledge that this approach builds on historic cost-benefit analysis but tries to go further. They challenge practitioners and experts to deconstruct the model, and improve and refine it.

SROI requires the ability to identify and measure a range of hard and soft outcomes. It is also dependent on using a participative evaluation approach in order to generate an informed understanding of what needs to be measured – particularly the unintended consequences of programmes which may generate significant public savings.

The New Economics Foundation calls this approach "narrow" SROI. We are currently looking for ways to develop a "broad" SROI which enables organisations to create a non-monetary way of relating financial and social

outcomes in order to capture the full range of impacts and not just those that can be easily attributed to changes in public expenditure. Examples of such broader measures include user satisfaction, social capital or trust.

SROI not only enables a fuller assessment of value for money. It can also be used to demonstrate the importance and impacts of those organisations that may create little economic value but whose social value is extremely great. This is particularly important for those activities which may require some form of subsidy since they cannot be fully self-financed. The analysis gives the public sector or other funders a sense of the level of subsidy required in order to generate a certain level of social return. In some cases that subsidy would in effect be an investment since long-term monetary gains should ensue.

SROI could also be important for enabling managers to track projects and support funding applications. It also enables policy makers to have a clear view of the impacts of funding.

The approach is still being developed since it has a range of difficulties. For example, it is hard to compare what may appear to be similar projects since they often use slightly different approaches and focus on different target groups. There is currently some experimentation with weighting of different returns. For example, the impacts of helping people who have been unemployed for 6 months into work might be given a lower weight than for 2-year unemployed people. However, such approaches can be quite subjective. It is also often difficult to correctly identify and quantify public cost savings. Care has likewise to be taken with a low index since this may mean that an organisation could in fact be generating high social impacts which are difficult or impossible to monetise. There is also a problem with attributing cause and effect and avoiding double counting. NEF is currently doing work to address these issues and further develop the model.[20]

This type of analysis can potentially support an approach to funding which is more about investment than about grant-giving and project support. Funders are paying for value created and not just compensating for costs incurred. SROI further promotes outcome and not output funding. It also has implications for internal decision-making and for funders' decisions as well as potentially creating new forms of social investment financing or social payment systems where funding moves away from short-term grant regimes to outcome-oriented incentive schemes.

Conclusions

This paper has highlighted the challenges and opportunities for the evaluation of third-sector impacts on local economic development. Policy-makers need to increase their understanding of these impacts and better support the third-sector in measuring outcomes. Whilst it is easier to evaluate

top-down programmes and create validation, or otherwise, for public spending on this basis, it is clear that these are not the only solutions to local economic development, particularly in highly disadvantaged areas. It is in reality the range of different activities and actions undertaken in an area, including by the third-sector, and the synergy between them, which will create long-term change. Many of those impacts will be intangible. Evaluating this messy reality is difficult but necessary to ensure that public money is effectively used rather than, as this paper has suggested, all too easily leaking away.

Notes

1. This is not arguing for targeted procurement necessarily, but for including a broader range of potential outcomes that can be delivered by any provider.
2. See for example, Westall A. (2001), *Value-Led Market-Driven: Social enterprise solutions to public policy goals*, Institute for Public Policy Research.
3. See Walker P., Lewis J., Lingayah S. and Sommer F. (2000), *Prove It: measuring the effects of neighbourhood renewal on local people*, NEF, *www.neweconomics.org*.
4. See for example, Wainwright S. (2002), *Measuring impact: A guide to resources*, National Council for Voluntary Organisations (NCVO), UK.
5. Found in a study by Price Waterhouse for the then Department of the Environment (1995): *Tenants in Control: An evaluation of tenant-led housing management organisations.*
6. Platt S. and Treneman A., The Feel Good Factor: a citizens' handbook for improving your quality of life, Channel 4 Television, London, 1997.
7. Morley E., Vinson E. and Hatry H. (2001), *Outcome measurement in nonprofit organisations: Current practices and recommendations*, Independent Sector, *www.independentsector.org*.
8. Bozzo S.L., and Hall M.H. (1999), *A review of evaluation resources for nonprofit organisations*, Canadian Centre for Philanthropy.
9. Kendall .J and Knapp M. (1999), *Measuring the performance of voluntary organisation activities*, Belfast Voluntary Activity Unit.
10. Morrisey, M. and McGinn, P. (Community Evaluation Northern Ireland) (2001), *Evaluating community-based and voluntary activity in Northern Ireland*, The Voluntary Activity Unit, DSD.
11. This approach can be found on the Scottish Community Development Centre website: *www.scdc.org.uk*. Achieving Better Community Development (ABCD).
12. This manual is available on the New Economics Foundation website: *www.neweconomics.org* or in hard copy. Walker P., Lewis J., Lingayah S. and Sommer F. (2000), *Prove It: measuring the effects of neighbourhood renewal on local people*. The model is being developed further and information will be available on the website.

13. Dewson S. *et al.*, (2000), Guide to Measuring Soft Outcomes and Distance Travelled, The Institute for Employment Studies, *www.esfnews.org.uk/evaluation/documents/distance.pdf*.

14. Dewson S., Eccles J., Tackey N.D., and Jackson A. (2000), *Measuring Soft Outcomes and Distance Travelled: A Review of Current Practice*, DfEE Research Report RR219.

15. The website of community indicators can be found at *www.rprogress.org*. The New Economics Foundation publication is *Communities Count! A step by step guide to develop community-based sustainable indicators*. This manual can be accessed through the website at *www.neweconomics.org*.

16. See *www.cbs-network.org.uk* and *www.socialenterprise.co.uk* for examples and toolkits.

17. See *www.ethicalexplorer.org* for further details. Ethical Explorer will be launched in autumn 2003.

18. See *www.redf.org* for details of OASIS and the Rubicon example below.

19. See *www.redf.org*. The paper which outlines the SROI metrics is the *SROI Methodology Paper*. There are other papers and resources relating to SROI on this website.

20. Our initial conclusions suggest that a stakeholder analysis is critical followed by an input/output and outcome analysis. A cause-effect model then needs to be created. There is a trade-off between data accuracy and data collection costs. National averages can be used instead of actual data but there is currently no accessible resource of such data.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 14

Methodological and Practical Issues for the Evaluation of Territorial Pacts The Experience of Italy*

by
Paola Casavola,
Department for Development – Evaluation Unit,
Ministry of Economy and Finance
Rome, Italy

* I wish to thank Francesca Utili for our common research on the subject and Laura Tagle for the many discussions on evaluation of local development issues. Responsibility for the opinions expressed in this paper rests with the author only.

Italian territorial pacts

A territorial pact is a specific policy instrument aimed at promoting local development through financial incentives to a group of locally based and integrated projects designed by a coalition of local actors (private and public). In some cases financial resources for technical assistance are also made available to the coalition.[1]

In Italy, as of September 2002, the number of territorial pacts approved and considered eligible for financial support is quite large. There are 230 pacts, of which 220 (including 91 pacts specialised in agriculture and fisheries) were selected through a national procedure, and 10 selected during the update of the programming of the Community Support Framework for OB.1 1994-1999 on the basis of a procedure agreed with the European Commission. The two procedures differ. In the national procedure the selection of pacts projects (initiated by a public national bid) simultaneously ends in the approval of the general pact project and of all the single initiatives included. The national procedure has been repeated and modified over time, so we now have different cohorts of national territorial pacts approved. The European procedure was only implemented once. It was carried out in two stages. First, 10 general projects and territories were chosen and later – through territorial bids – single initiatives were selected.[2]

Although single territorial pacts do not cover very large areas, the instrument is so widespread that a very large portion of the national territory is affected (see Figure 14.1).

Territorial pacts in Italy have a very mixed reputation. Among politicians, general observers and territorial experts they have fierce enemies and determined defenders. Most of the debate (and most of the arguments in favour or against) have not however been based on scientific evidence, rigorous monitoring or evaluation research. They have rested mainly on conjectural arguments, direct experience and, often, prejudices. In what follows, however, we do not consider this general – and mainly media-driven – debate in detail. Rather, we concentrate on what and why we might want to learn from evaluating territorial pacts and report on what has been done so far in this respect.

Figure 14.1. **Italy: areas covered by territorial pacts* (September 2002)**

1. *The 91 pacts specialised in agriculture are not included (as they mainly overlap with other pact areas).

Objectives of territorial pacts and evaluation questions

The territorial pact is built around the idea that financing a coalition of actors with a project could serve the purpose of bringing individual agents together, giving rise to economies of agglomeration and capable partnerships. In theory, if the desired outcome is to promote economies of agglomeration and collective capacity, this line of action should be superior to financing the single worthy but uncoordinated projects of individual private actors or public entities (as it is the case for common incentives for private investments, or the financing of single public initiatives such as infrastructure, training, or communal and social services).

The idea is not new in local development promotion. It is based on both theoretical and empirical findings showing that areas where development has been spurred often possess dense social and economic interrelations. These interrelations might appear as formal, informal, market-driven or institutional. Linkages in production activities and related services also match these relations among the relevant local actors. Local economies based on coherent agglomerations of a variety of activities have often proved to be associated with local prosperity, comparable to – or even more long-lasting than – forms of territorial development coming from the presence of single large firms and plants. This is, for instance, the case of the so-called Italian industrial districts. Economists and social scientists have studied the latter extensively. However, these studies concentrate on natural evolution and equilibria, hence their findings do not necessarily support the idea that it is possible for a policy maker to promote or accelerate local development by devising incentives for a coalition to form or progress more speedily. The need for evaluation comes in part from this last consideration.[3] The general evaluation question in the background of this line of reasoning is whether, to what extent and in which circumstances the policy maker can induce or accelerate local development dynamics.

In order to identify the specific relevant evaluation questions concerning territorial pacts, it is useful to consider the explicit objectives of the policy. The policy has two specific ambitions seen as crucial mechanisms for inducing or enhancing local development. The first is to support the start-up of a sound, locally-rooted, integrated project made up of a set of different initiatives involving responsibilities on the part of many public and private actors. If the project is successful, it is expected to generate positive spillovers for the economy of the area. The second ambition is to promote the formation of a robust local coalition, a group of actors that might – through the setting and implementation of the original project – learn how to interact with each other and promote further development. This second objective is a key one, as permanent changes are associated with the creation of a long-lasting local coalition.

These two objectives of territorial pacts call for two sets of evaluation questions. The first set relates to the nature of the integrated project financed (for example, under which circumstances has the policy mechanism proved successful in inducing – and selecting for financial support – a good project? Is the project well defined and rooted in a real knowledge of the potential of a territory? Is it capable of triggering a local process of development? And how? etc.). A somewhat different set of questions is related to the nature of the local coalition that the pact is to promote (is it a good coalition? Is the process helping in inducing or enhancing fiduciary relations among actors? Has their collective capacity in problem solving been augmented? In which way? etc.).

The questions sketched above are related to the specific mechanisms of development that the pact should directly activate (good projects and institutional capacity). However most policy makers are mainly interested in the final economic results for the territories where pacts are implemented. In other words, there is a third and more explicit set of questions of the following type: have the pacts promoted local development, firms' growth, employment opportunities? These latter questions are indeed important, but they can also be misleading if not addressed properly. Two issues must in fact be considered: the time span between implementation of the original project and the desired spill-over effects; the circumstance that questions of this type are indeed very general. In discussing evaluation of territorial pacts in Italy this latter point is particularly important as pacts are implemented in territories (especially the Italian Mezzogiorno) in which other policies are at work at the same time and also rely on different mechanisms. The need to disentangle effects coming from -or cumulating by – different kinds of interventions is of primary importance in discussing evaluation methods. Hence we need to keep this in mind when we look at territorial pacts in order to make clear what we expect from evaluation exercises.

The cultural and institutional environment for carrying out evaluation exercises: theory and practice

As in other contexts, we can evaluate both for learning (how the instrument actually works; where it works well, where it doesn't and why) and accountability[4] (to the fund givers). In the Italian case, however, most national policies are not evaluated *ex post* (or an on-going basis) and systematic policy monitoring has only recently begun. To some extent this is a good opportunity for territorial pacts. In fact, it opens a window of opportunity for promoting a learning approach, which is both less threatening for the policy maker and particularly advisable when the policy in question does not have an evaluation history.

In evaluating pact performance a crucial issue is timing. In other words, given the characteristics of the instrument and its mechanisms, it is important to be both patient and forward looking. First, for assessing final results (both in terms of local development and institution building) it is necessary to wait for the pacts to accumulate enough implementation history. However, in order to be able to assess results it is necessary to collect a certain amount of information while the pacts are under implementation and to prepare the necessary statistical information at the desired level of territorial detail. These arguments were perceived as very abstract only a couple of years ago. However, both these concepts have now been fairly well understood by many policy makers.

Despite the presence of a certain residual degree of impatience,[5] the administration has decided in April 2002 to launch a study involving fieldwork on a group of pacts that have been in place for several years (signalling awareness that only in this case can a study come up with some answers about performance). Moreover, a new monitoring system has been set up which should be able to provide precise information on the administrative history of financial contributions to the pacts and offer other information which in principle could be used in conjunction with other territorial data. To implement a real evaluation, however, it is necessary that the policy maker express an entire and detailed set of evaluation questions in which s/he is genuinely interested. If there is an interest in learning from evaluation, these questions should be expressed not only in terms of socio-economic results occurring in the territories, but also in terms of the functioning of the instrument itself (does the set-up procedure to gather and select the projects work? what are the flaws? what are the good and the bad incentives?).[6] In this respect, more progress is needed. Even in the study recently launched, despite an indisputable genuine interest about what happened in the different contexts of the pacts, not very much time has been spent in detailing questions challenging the role of the administration in designing and implementing the instrument.

Another key issue is related to evaluation research methods. Methodology is crucial for at least two reasons. First, as a territorial pact is a package of different things (it is not a single well defined policy intervention, but an entire set of different instruments pooled together), in order to learn from evaluation we need to know about key ingredients (which aspects of the policy are the most important/effective) and the methods used have to be appropriate to this scope. Second, even more than in other policy interventions, context matters. The evidence coming from administrative monitoring of financial contributions to single initiatives included in the pacts makes clear that they proceed very differently. There appear to be different mechanisms at work. We want to know why some pacts seem to work better

that others, to what extent this is true and why this happens. The methodology used has to deliver answers that help in uncovering reasons for differentiated outcomes. In other words, even more than in other circumstances, we need a method of investigation which is able to give explanations. All these considerations call for a methodological approach which considers primarily comparative analysis between pacts (coming from accurate fieldwork to be read in conjunction with more macro statistical evidence). In this respect, even if the study recently commissioned is not strictly speaking an evaluation study, the methodology chosen (fieldwork and direct interaction with relevant stakeholders in territories where the pacts are active) appears to be adequate enough to give some revealing answers.

Another piece of good news for evaluation is the increasing institutional attention devoted to building territorial statistics at a very detailed administrative and economic level ("local systems" and council level). This kind of information has important implications for evaluation of territorial pacts – or other similar forms of local development promotion – as it allows the assessment of changes in target areas using statistical concepts similar and thus comparable to those available for larger territories (at regional, or province level). Of course, as suggested before, the availability of more detailed territorial statistics does not imply that evaluation of the effects of territorial pacts can be limited to comparing areas where the instrument is active and areas where the instrument is not active.[7] However, standardized detailed information allows for several kinds of informative analyses that can be used in conjunction with other methods of investigation and that also offer a possible guide for picking areas of interest (for instance, areas where target case studies might be carried out on specific instruments (like territorial pacts), or also other very informative kinds of evaluations, as for instance area–based evaluations).[8]

What we have learned so far: academic studies, preliminary evaluations, direct experience and learning by monitoring

Most of the studies carried out in the last few years on territorial pacts have involved independent academic research, mostly fieldwork by sociologists who focussed on the issue of social capital accumulation.[9] The nature of the coalitions, and the development of trust relations has been the object of many of these studies. As these are mainly case studies undertaken with different methods, they were not aiming to reach general conclusions. However, they all seem to suggest that the process associated with the pact, in the various cases examined, did trigger some social capital dynamic, even if in a differentiated way. In the academic debate on these issues these are important findings that speak loudly in favour of the possibility that pacts may affect local development. Nevertheless these studies also seem to suggest

that if the pact offered an opportunity, it took some pre-existent local social leader to glue the coalition together, an issue that should be studied more extensively.

Sociologists also found different models at work in the coalitions' formation, some of which are in fact quite perverse (they observe in some instances the formation of opportunistic coalitions). As we do not have these kinds of studies available for a sufficiently large number of pacts, we do not know whether good coalitions outnumber opportunistic ones. However it is clear that this variety of outcomes shows that the procedure used to select the pacts had some limitations in discouraging opportunistic coalitions. It would be useful to investigate whether the magnitude of the financial incentives and the mechanism for selecting the projects played any role.

In the first part of 2000 a survey was carried out on all the pacts in operation (46). This survey was based on a structured questionnaire for entrepreneurs and key informants.[10] The survey was part of an autonomous study aimed at uncovering the motivation of actors who joined the pact coalition, learning directly from involved entrepreneurs about the opportunities and needs of the territories. The main result of the study was that the policy was particularly well received in most of the cases (entrepreneurs were on average quite happy), despite some complaints about lengthy procedures.

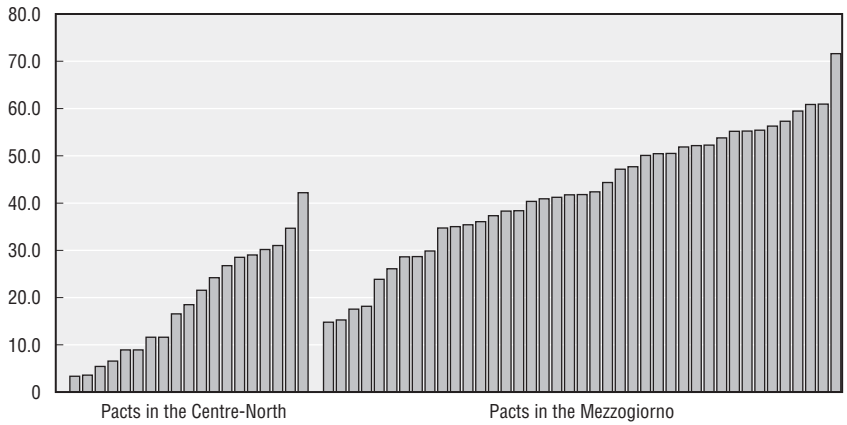
What is most important, however, is that the study produced a database with coded information on the agents' perceptions and point of views on a set of issues.[11] In the same year, within the Department for development (the administrative authority in charge at the time for implementing the pacts) another project was carried out to build economic statistics defined at the pact level, in other words referring to the territory included in the pact. That exercise showed that pacts were starting in territories with different economic conditions, though the very first cohort of pacts appeared in areas that were relatively less disadvantaged.

The most thought provoking finding came, almost casually, from monitoring evidence, as early as the end of 2000.[12] Up to that point the general debate on territorial pacts had considered pacts as homogeneous (pacts were thought and portrayed as being everywhere the same, suffering from similar weaknesses that were mostly due to bureaucratic inefficiencies). Monitoring showed that pacts were behaving differently with respect to the progress in the pact project (see Figure 14.2).

Some pacts, after controlling for their starting time, showed a significantly higher rate of active initiatives and a significantly higher rate of expenditure. The issue seemed worth further investigation.

Figure 14.2. **Evidence from monitoring the implementation of the projects in the first 61 approved territorial pacts**

(Centre North and Mezzogiorno; percentage of financial contribution used by each pact, standardized data referring to April 2001)



We then tried to investigate whether and to what extent the initial conditions of the territories, as measured by a simple set of variables, had affected the formation of the local partnership, its objectives and its functioning.[13] In particular we were interested in uncovering whether those partnerships which seemed to perform better (as they appeared to progress more quickly in the implementation of the pact's project) could be predicted by more favourable initial economic conditions in the territories or, instead, whether they had acquired their skills in the process of building and implementing the pact. The main idea was to look at differences in ongoing performance in order to uncover general characteristics of the instrument at work, asking explicit evaluation questions (where a pact works better, why? Is this because of things that the policy cannot affect – like more favourable initial conditions? Is this because of something that can be incorporated in a better policy design?)

The study used statistical techniques exploiting a quite rich data set built on the basis of administrative monitoring information about the pacts' relative realization speed (assumed as a proxy for intermediate performance), characteristics of the pacts' areas and data coming from the 2000 survey.

The exercise was carried out over the first 61 pacts, with a specific focus on those implemented through the national procedure, for which the performance variable was measured more accurately. The results of the analysis can be summarized as follows. Relative initial economic conditions did not seem to affect the implementation of the pact (see Table 14.1).[14]

Table 14.1. **Pacts' implementation and initial conditions of the territories**

PERF	A		A1	
Variables	Coefficient	(t)	Coefficient	(t)
Constant	3.301	(0.752)	2.012	(0.494)
DC1	0.547	(1.725)	0.312	(1.036)
DC4	0.455	(1.327)	0.289	(0.900)
LVALAG	–1.242	(–2.286)**	–0.652	(–1.221)
LUNR	0.265	(0.745)	–0.241	(–0.662)
LSIZEFC	0.99	(0.517)	0.092	(0.523)
DS			1.186	(3.242)**
	F (5.55) = 6.04 Prob > F = 0.0002 Adj R2 = 0.2957 Num Obs 64		F (6.54) = 7.65 Prob > F = 0.0000 Adj R2 = 0.3995 Num Obs 61	

PERFi = CONSTANT + b1 (DC1) + b2 (DC4) + b3 (LVALAGi) +b4 (LUNRi) + b5 (LSIZEFCi). PERF is a variable proxying for the ongoing performance of the pact at April 2001 (see explanation in the text); DC1 is a dummy for the first cohort of National territorial pacts which were approved much earlier; DC4 is a dummy identifying territorial pacts approved with the European procedure; LVALAG is the log of per-capita value added in the area of each Pact in 1996; LUNR is the log of the unemployment rate in the area of each pact in year 1996 (chosen as the initial conditions date); LSIZECF is a control for the (log) dimension of the total public contribution available for each pact. The same regression is then run adding a dummy DS for location in the South. OLS estimates, bounded variable are treated with a logistic transformation.

Source: Casavola, P and Utili, F.(2002) Promozione di partnership locali per incoraggiare lo sviluppo locale: una valutazione preliminare dei patti territoriali, Sviluppo Locale, Vol. IX, n. 20.

This turned out to be a quite robust result holding among the various subgroups of pacts examined. If initial economic conditions do not count very much in explaining differentiated performance, we might conclude that it is possible to foster local development through incentives to form coalitions even in areas that are very disadvantaged.

Instead, the performance of the pacts [as supported by an exercise carried out on a smaller number of pacts for which additional information was available (see Table 14.2)] seemed to be particularly related to the quality of the process of setting up the pacts at the local level.[15] Pacts in which relevant actors had been involved in discussing the project from the very beginning appeared to proceed faster in their implementation, a result confirmed by other fieldwork.[16] The result is interesting, as the way the local process is organized is a variable that can be directly affected by the mode in which the policy is delivered in a territory. It signals the importance of giving enough time and resources to the initial phase when designing a policy for promoting the formation of a local coalition. It also suggests that accurately screening the process of coalition formation (in order to prevent weak projects) could be crucial for their future success. This result is also particularly informative as it can be generalized to a family of policy instruments based on a bottom-up approach to development.

Table 14.2. **Pacts' implementation and the quality of the local setting-up process**

PERF	B1		C	
Variables	Coefficient	(t)	Coefficient	(t)
Constant	–5.3035	(–1.37)	–5.058	(–1.326)
DC1	0.228	(0.815)	0.206	(0.737)
LVALAG	0.479	(0.77)**	–0.054	(0.896)
LSIZEFC	0.097	(0.483)	0.054	(0.272)
DS	1.236	(3.417)***	1.186	(3.294)***
LPARTIC			0.279	(2.195)**
TP_TRUST			–0.99	(–1.014)
TP_INFR			–0.076	(–0.674)
	F (4.31) = 6.02 Prob > F = 0.0011 Adj R2 = 0.364 Num Obs 36		F (7.28) = 4.35 Prob > F = 0.0023 Adj R2 = 0.401 Num Obs 36	

PERF= CONSTANT + b1 (DC1) + b2 (LVALAG) + b3 (LSIZEFC) + b5 (DS) + b6 (LPARTIC) + b7(TP_TRUST) + b8 (TP_INFR). In the equation the new variables are all derived from a survey carried out on a smaller number of pacts, asking questions to entrepreneurs taking part in the pacts: LPARTIC is a variable derived from positive answers to a question in which entrepreneurs were asked whether in the initial process of setting up the pact all the key local actors had been involved; TP_TRUST is a variable derived from answers to a question in which entrepreneurs were asked about the role of the pact in inducing, enhancing or hampering trust relations among local actors. It represents the share of entrepreneurs that perceived a positive role of the pact in this respect; TP_INFR is derived from another set of answers to questions related to the most urgent necessities of the territory and it represents the share of entrepreneurs who signaled the need for better infrastructures.

To learn more however it appears necessary to carry on more targeted fieldwork. This is the task of the study implemented by a group of researchers working for the department of Development Policies. In particular, the researchers have been asked to come up with a more appropriate definition of success and investigate more deeply what makes the instrument work better.[17]

Notes

1. Resources for technical assistance are provided for preparation of the projects and during implementation to support the coordination process.
2. There are other relevant differences among the two groups of pacts, the main ones being the greater role in European pacts played by the local agency responsible for coordinating implementation (a feature that in time should be extended to the national pacts as well) and the presence of a much larger variety of initiatives (whereas in the national pacts the public contribution is given mainly to finance entrepreneurial activities and to a much lesser degree public infrastructure).
3. The industrial districts literature has stressed the importance of context and local relations as a basis for different paths to development and this argument apparently speaks in favor of policy instruments like territorial pacts. However,

the literature on social capital inspired by other work (following Putnam's famous contributions on the origin of Italian regional development) has often reached the conclusion that it takes a long historical path to build social capital. Some researchers have hence inferred that there is no room for the policy maker in trying to speed up this process.

4. I am using the term accountability in a somewhat improper and restrictive sense for the sake of the argument and just to keep in mind that some of the questions usually raised in relation to territorial pacts are of the following sort: are we spending all the money well? Overall, is the instrument providing the promised increase in development?

5. In the recent political debate some commentators, understandably concerned with the large number of existing pacts, have asked for a quick decision on which (of the many in place) are the good pacts – to keep supporting – and which are the bad pacts – to possibly close down.

6. In other words an important set of questions relates to the relation between the policy design and the policy outcomes. In fact the general characteristics of the policy (financing a coalition with a project) do not suffice to completely describe the nature of the intervention. A series of other design features should be considered: what kind of projects and with which characteristics; which actors are eligible to present the project; how large is the public contribution offered; how and when is the project submitted; which procedures should be followed to select projects to finance. All these elements might play a role in determining at least in part the future success of the coalition selected.

7. There are numerous theoretical reasons to discourage a mechanical application of the so called standard evaluation paradigm (which compares a treated group to an untreated group) to instruments of local development like pacts.

8. Area-based evaluations could be a promising method for evaluating local development policies as investigators are required to start their work looking at a limited portion of the territory and from that angle reconstruct the possible causal chains linking observed facts and behaviors to policy interventions in that area.

9. See for instance the 2001/3 issue of the Italian journal Stato e Mercato, almost entirely dedicated to territorial pacts.

10. Sviluppo Italia – Iter (2000) "Caratteristiche e potenzialità dei Patti territoriali", Roma.

11. Most of the sociological studies quoted above had not produced coded or standardized information (to be used in other studies).

12. Ministero del Tesoro, Bilancio e Programmazione Economica, "Terzo Rapporto sullo sviluppo Territoriale", November 2000.

13. The results reported briefly here come from Casavola, P – Utili, F. *Promozione di partnership locali per incoraggiare lo sviluppo locale: una valutazione preliminare dei patti territoriali*, Sviluppo Locale, Vol. IX, No. 20, 2002.

14. We looked at the relation between a dependent variable describing the ongoing performance of the pact at some point in time (the amount of public financial contribution actually used at April 2001 over the total amount available at the same date, on the hypothesis that all the initiatives were progressing at the fastest rate) and indicators proxying the initial economic condition of the area of the pact (per capita value added and unemployment rate computed for a period before the activation of the pact). A few other controls were also added.

15. The variable was recovered by the 2000 survey and represents the degree of completeness; as perceived by entrepreneurs in the pacts; of the set of local actors involved in discussing the pact project.

16. Dipartimento per le politiche di sviluppo e coesione, Relazione sui patti Territoriali (June 2001).

17. The study was carried out in the second half of 2002 and involved field work on 19 cases. The final report came out in January 2003. Results from the study are not easy to summarize without losing interesting details. However, the researchers did come up with an interesting list of items that can be associated with success (spanning from the capacity of the coalitions to provide innovative projects to their ability to find new sources of finance to implement other projects not initially included). A synthesis of the research is provided on-line at *www.dps.tesoro.it/documentazione/docs/patti/RICERCA_PATTI_TERRITORIALI.pdf*.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 15

Evaluating Territorial Employment Pacts – Methodological and Practical Issues The experience of Austria

by
*Peter Huber**
Austrian Institute of Economic Research (WIFO),
Vienna, Austria

* The author would like to thank Hedwig Lutz for helpful comments, errors remain the author's responsibility.

Introduction

Territorial employment pacts are interventions in local governance regimes aimed at generating institutions and social capital in regions to improve economic policy. The underlying premise of this intervention is that policy activities, which are developed autonomously, collectively and locally by encompassing partnerships are likely to contribute to improved effectiveness and efficiency of policy delivery. Clearly, an evaluation of territorial employment pacts has to take an empirical view on this premise. In the best of all cases an evaluation would establish whether, in what way and to what extent individuals in a region have profited from territorial employment pacts. It individuals have profited, it would then attempt to compare benefits to the costs incurred and derive a net benefit of the pact.

Such an evaluation, however, would be burdened with methodological problems. In particular, in labour market policy evaluation, any success of a particular measure is usually attributed to the measure rather than to the institution that designs it. Evaluating the efficiency of institutions in implementing programs implies a different counterfactual from that in much of the active labour market policy evaluation literature. For example rather than asking, "what would have happened to a particular person if he/she had not been included?", the counterfactual here should help to answer the question, "what would have happened if the institution designing the measure had not existed?". This counterfactual may be very difficult to identify precisely.

This paper is concerned with outlining the evaluation approach chosen to evaluate territorial employment pacts in the framework of the ESF Objective Three Evaluation in Austria. In particular, the paper argues that using concepts stemming from process evaluations, approaches may be able both to identify the achievements of pacts as well as the impediments to their success. The next section of the paper describes some features of Austrian TEPs relevant to evaluation. Subsequent sections briefly discuss data issues and outline the evaluation method chosen. A final section concludes.

Particularities of territorial employment pacts in Austria

Although there is a long-standing tradition of regional labour market policy co-ordination in Austria, territorial employment pacts are a relatively new policy instrument. The original initiative came from the European

Commission. In 1997 it called for submission of projects under an initiative to improve the employment situation. The intention of this measure was to combat unemployment through the design of specific programs under the title "territorial employment pacts". Four Austrian pacts (Salzburg, Tyrol, Vorarlberg and Vienna) were selected by the Commission for this initiative.

The idea of territorial employment pacts was well received and in 1998 the former Ministry for Labour, Health and Social Affairs encouraged the regional offices of the public employment service (AMS) as well as the provincial (Bundesländer) governments to conclude further TEPs. In the framework of the national action plan for employment special subsidies were introduced to support such territorial agreements.

Based on this national initiative TEPs developed rapidly. By 2002 a provincial TEP was established in each of the nine provinces in Austria (there are thus currently nine provincial TEPs in Austria.) Furthermore, based on the recommendations of an early OECD study (Campell, 2001) a number of pacts had devolved their initiatives to a lower regional level through various institutional arrangements.

The provincial pacts, which are the primary focus of this paper, vary widely in their goals and how they define their role. Most pacts (*e.g.* Upper Austria, Vienna and Lower Austria) put particular emphasis on their role as a forum to co-ordinate policies (in particular active labour market policy budgets of provincial PES and economic policies of provincial governments) both in terms of budgetary co-ordination and policy design, putting less emphasis on their role in designing innovative measures. Some pacts (*e.g.* Styria), however, define themselves as a pool of innovation responsible for the design of new measures to improve co-ordination with economic policies in their province. Pacts which define their primary role as a co-ordination instrument as a rule take responsibilities for co-ordinating the use of budgets of both the provincial public employment services and provincial governments although these funds are not actually administered by pacts.[1] In these pacts the issue is thus what has been the "value added" of pacts in co-ordinating these budgets. Furthermore, the extent to which this co-ordination extends beyond budgetary co-ordination, to a general discussion of relevant policies, varies among pacts.

A further particularity of Austrian TEPs is that they operate at the provincial level, that is on relatively large regional units (on average around 1 million inhabitants) which are often characterised by internal differences in labour market conditions. As pointed out in an early analysis by the OECD (see Campell, 2001) this provides pacts with the necessary resources and ensures the involvement of actors with substantial decision making powers. But the pacts may also be too large to ensure the involvement and motivation of all potential decision-makers.[2]

Biffl *et al.* (2000) argue that in the majority of cases TEPs were founded as bodies for co-ordinating the activities of the provincial AMS and provincial governments. These two institutions were the dominant partners in most TEPs. Also, according to Biffl *et al.* (2000), there was a strong focus on public sector institutions and on the social partners' involvement, rather than private sector institutions in the partnership.[3]

Despite their substantial heterogeneity, there are a number of official documents published both by the ministry as well as by the co-ordination office for the Austrian pacts, which is responsible for the co-ordination of territorial employment pacts on a federal level. Although these documents in general tend to be relatively imprecise concerning the concrete problems to be addressed by pacts, the documents highlight the role of pacts in co-ordinating economic and labour market policy.

Furthermore, the co-ordination office presented a list of common features for pacts. In particular, this list suggests that all pacts must be based on an analysis of the existing labour market problems, strategies and goals shared by all actors.[4]

Although pacts are relatively new in Austria, they have repeatedly shown interest in external evaluation and in consulting on future development. Early studies mostly discuss the structure and development of pacts. Campbell (2000) for instance suggested that with respect to the optimal regional scope of TEPs an analytical differentiation should be made between the strategic and operative levels of pacts. For the strategic aspects of pacts, it is imperative to involve decision-makers who have the relevant decision-making powers and a command of adequate resources. This suggests organisation at a larger regional level. The operative level of pacts, by contrast, has to secure the involvement, participation and motivation of all relevant local actors. This can best be done in a smaller regional context, resembling that of NUTS3 level regions. This suggestion was followed by most pacts in the larger provinces of Austria.

More recently, Leitner *et al.* (2002) have conducted a detailed and careful evaluation of the impact of the territorial employment pact of Vienna, and an evaluation of the Styrian Pact is in progress. Leitner *et al.*'s most important findings are that measures of the TEP are more intensive and more targeted than other measures, and that these measure have created net job gains. However, they are critical of the fact that relative to TEPs in other countries the Viennese TEP is still not focused enough.

Furthermore, the national co-ordination office (ZSI) has created a number of measures aimed at self-evaluation. For example, interviews were conducted with the presidents of the social partner organisations (see: Scoppetta, 1999 and ZSI, 1999a). While statements are not uncritical of, for

instance, the lack of initiatives for encouraging entrepreneurship (in the case of employers) or the need to create more local-level operative initiatives (trade unions), and while they also draw attention to the need to evaluate pacts from both an individual and comparative perspective, all social partners welcome the idea of TEPs and highlight their importance for local policy co-ordination. In particular, social partners praise the role of TEPs in raising awareness of the need of policy co-ordination, their successes in motivating a wider range of regional actors and their ability to contribute to increased flexibility in labour market policy.

Data situation concerning pacts

One particularity of territorial employment pacts is that they generate only few administrative data. In some pacts, which are involved in co-ordinating policy measures rather than conducting or designing such measures, data on the budget provided to pact measures, persons involved in measures and so on are hard to obtain and have very little meaning. Furthermore they are not collected by pacts but by the partners. Pacts, however, do generate substantial amounts of information in the form of texts which they are either obliged to provide or provide on their own account. Pacts are required to provide the following documents:

- The contract of the pact – this contract is renewed regularly but can extend to a number of years. It contains information on the partners of the pact, an analysis of the existing labour market situation in the region, and details on the strategy proposed by the pacts.
- The working program of the pact – this contains additional information on the goals, strategies and analysis in pacts where the contract is not renewed annually.
- The request for subsidies from ESF – this provides details on the funds requested from the ESF and specifies for what purposes these funds are used, the partners to the pact (detailed by financing, supporting and consulting partners) and some information concerning the problems of the region and the strategies proposed by the pact.

Since some of these documents are submitted annually, or at least on a regular basis, they are able to provide substantial insights on both the current state as well as the development over time of the partnership, the goals set by the pact and the quality of the shared analysis of the pacts. Furthermore, a number of pacts have published additional information in the form of detailed monitoring and implementation reports, strategies and studies concerning either the further development of the pact (such as integration of regional structures, etc.) which are additional sources of information for evaluation.

These documents are provided by pacts and are as heterogeneous as the pacts themselves. Furthermore a standardised report on pact activities, partnerships and other details of the pact is provided annually by the national co-ordination office (ZSI) for all pacts.[5] In addition, sources such as published economic policy documents of the provincial governments as well as the documents of the PESs exist in all provinces.

This data situation suggests that an evaluation of territorial employment pacts should – aside from being based on interviews both among pact partners as well as the persons responsible for pact management – take due account of these sources of information to generate objective indicators on pact development.

An approach to evaluation

An evaluation of pacts must thus take into consideration the existing data situation as well as the institutional arrangements, such as the extent and quality of the partnership, the way goals are defined and the way in which they are implemented as part of the evaluation process. In particular such an evaluation has to take into account that:

1. TEPs are new partnerships, which are strongly oriented towards changing the behaviour of regional actors in such a way as to provide for co-ordination of different policy fields on the level of analysis, strategy and implementation. In this context, the "quality of the partnership" is of central importance to the long-run success of the measure.
2. TEPs are in a continuous state of development and learning. Thus these learning processes must be taken into consideration by the evaluation, in order to understand that mistakes made, when corrected, may have been important preconditions for learning.
3. TEPs are only one of many institutions operating in the implementation of labour market and employment policy. Thus their relationships to other institutions must be considered to get a full picture of the value added of pacts. This is particularly important since multiplier effects as well as substitution and displacement effects between institutions could arise.

In principle, an evaluation in this context could choose to focus on a number of aspects of TEPs. For instance one could choose to focus on the pact as an institution. In this case particular emphasis would be put on the role of the actors in the partnership, the nature and extent of co-operation, its goals and its learning processes. The ultimate goal of this evaluation would then be to assess whether the pacts have contributed to establishing social capital in the region. The strength of this approach is that it is directly geared towards identifying whether one of the primary goals of pacts (namely to construct social capital in the region) was achieved.

Alternatively, the focus could be put on the pact as a program. In this approach the emphasis would shift to answering the questions of whether the strategy of the pact seems adequate and what its likely results will be.

Another option is to focus directly on processes (see Schmidt, 1996). In this approach the goal of evaluation is to analyse policy formulation, implementation and uptake as well as the effects of the policy in order to identify the connection between these elements. The advantage of this is that it encompasses the complete policy cycle to which territorial employment pacts are subjected.

Strategy formulation

At the level of strategy formulation the key issue is to determine how the formulated strategy is developed and who has influence over the formulated strategy. Thus the analysis of strategy formulation should take account of three separate issues:

- First, in the context of territorial employment pacts the presumption that social capital and a culture of co-operation among the actors lead to improved policy outcomes suggests that the "quality of the partnership" should be made a central part of the analysis. This "quality" of the partnership can be operationalised by a number of quantitative and qualitative indicators such as the extent of the partnership (number of partners, types of partners) taken from the pact documents, the presence of a "culture of co-operation" among the partners, the openness of the pact to outsiders, which can be gleaned from interviews with both outsiders and insiders, and the dominance of certain partners, which can be established from a combination of indicators such as location of the pact office, reflection of the partners goals in pact documents, etc.
- Second, an important element in the analysis of strategy formulation is to determine to what degree pacts have actually integrated policy fields at a strategic level. Based on the pact documents (defined in the last section) a number of indicators can be constructed concerning integration on the strategic level. In particular, one can assess to what degree the pact strategies address issues of more than one policy field and to what degree these policy fields are integrated into a single coherent strategy.[6] Furthermore, since a number of pacts have renewed their strategies over time, one can also check to what degree strategies have become more integrated. Also, by looking at published strategies of policy makers in various fields in a province, one can assess to what degree the integration has transcended the narrow scope of the pact documents, and found acceptance elsewhere. Another important aspect of this analysis is to look

at the content of the strategy developed by a pact and determine its relevance to regional problems as well as the potential for its success.

- Third the analysis has to take into account the changes in both the partnership as well as the content of the strategy over time. In this context indicators such as the number of changes of partners can provide important information on the stability and growth of the partnership, while looking at the development of strategies over time will provide information concerning the flexibility of the pact in terms of its strategies

Analysis of Implementation

In the case of implementation the focus is on the impediments to implementing the pact's programme or strategy. In particular the focus is on whether the pact is endowed with adequate organisatorial and financial resources in order to implement its strategy. In this context three steps of analysis are necessary:

- First, conflicts, which may impede the pact's ability to implement strategies have to be addressed. This is necessary because pacts are neither monolithic organisations nor do they operate in an institutional vacuum. Thus a number of conflicts could limit the ability of pacts to implement their strategy. These can be classified according to two dimensions: for conflict with other institutions or within pacts: and conflicts with institutions of the same regional level (horizontal) or another regional tier (vertical). Particular emphasis has to be given to the issue of whether conflicts arise because pacts have insufficient competencies to implement programs and to what degree this is the case. Information on these conflicts can, on the one hand, be collected in interviews. On the other hand information on the potential for improvement of the partnership as well as its problems can be provided from a structural analysis of regional actors and their relationships with each other.
- Second, in addressing implementation and conflicts among partners and with other institutions one has to ask, what would have happened if TEPs had never existed? and how has the creation of TEPs impacted on other institutions? Three effects of particular relevance can be analysed. These are: *a)* Dead-weight losses – one possibility is that some of the observed behaviour would also have occurred in the absence of the TEPs. In particular, the vagueness of the goals set for the TEPs seems to suggest a potential for such dead-weight effects. This may lead to situations where actors do not feel that the goals of the TEPs limit their actions, and thus subsidies are used to finance previously existing institutions;[7] *b)* Displacement and substitutions effects – these refer to the possibility that the presence of TEPs has limited the efficiency of other institutions with similar tasks. This may be of particular relevance in pacts

which see their main role in designing innovative measures, and which in consequence are in direct competition with other institutions designing such projects; and *c)* Multiplier Effects – finally, positive experiences with territorial employment pacts may have led to co-ordination over and above the co-ordination observed in territorial employment pacts. While there are a number of reasons to believe these effects may play a role for territorial employment pacts it is difficult to quantify them. To some degree, looking at the dynamics of the development, in particular in strategy formulation, may provide insight on the likelihood of dead-weight losses. Furthermore, in-depth interviews with pact partners, competitors and persons responsible for implementation may yield additional results.

- Third, aside from analysing the question of the interaction of individual pacts with each other, this step in the analysis has to take into consideration the endowment of pacts with both organisational and financial resources relative to their tasks, in order to assess the organisational efficiency of pacts. In this context the costs of setting up and operating TEPs have to be determined. Furthermore, information on financial resources available from the budget of the TEPs and additional information on costs of partners to the pact, in terms of time spent at meetings, etc., can be gathered from interviews among pact partners.

Policy take-up

Territorial employment pacts are designed to change the behaviour of regional actors in a particular fashion, which in a very general form could be specified as getting actors to: *a)* co-ordinate and communicate activities with each other; *b)* develop shared views on labour market policy problems; and *c)* design a coherent policy taking each other's actions into account. The focus of the analysis of policy take up is on establishing whether and in what way the behaviour of regional actors has indeed changed due to territorial employment pacts

To address this issue it is important to consider the incentive structure of the regional actors both in terms of extrinsic and intrinsic motives. In particular, issues of accountability (*i.e.* who gets the credit for the successes of pacts and who is responsible for failures) and of the transparency towards the outside are of primary importance in this analysis. While this issue has been shown to be of some importance in previous evaluation studies of territorial employment pacts (OECD, 2001), which find that the vague definition of "property rights" of the results of pacts leads to a lack of motivation among partners, data on this issue is obtainable from interviews with partners only.

Furthermore, a behaviourally based indicator of policy uptake can be formulated by observing the development of pacts themselves. If the policy

were taken up positively by the partners one would expect pacts to receive increasing competencies in the policy arena. Thus, observing whether pacts have deepened their regional activities, broadened the partnership or diversified their content can provide additional information on the up-take of pacts. Finally, successful policy up take also would imply that pacts comply with the common quality criteria established by the co-ordination office for Austrian pacts (see ZSI, 1999).

Conclusions

This paper is concerned with outlining the approach chosen to evaluate territorial employment pacts in the framework of ESF Objective three evaluation in Austria. In particular I argued that using concepts stemming from process evaluation approaches might be helpful to identify both the achievements as well as the impediments to the success of pacts and can provide insights into the further development of these institutions. While the proposed method thus goes some way in evaluating pacts, it is not free of problems. In particular, the approach proposed will encounter problems if too much is demanded of the evaluation in terms of quantitative estimates of labour market impacts.

Notes

1. Usually funds are co-ordinated by means of a contract between partners specifying the use of funds of a particular partner for different measures. The partners then administer the funds.
2. In Austria the provincial offices of the public employment service have the authority to design and implement appropriate measures for the territory of their respective province and to co-ordinate policy with provincial governments. District offices by contrast are only responsible for the implementation. Furthermore, provincial governments are responsible for developing economic strategies, spatial planning. They account for a substantial part of total expenditure. Below the provincial level communities are the only autonomous administrative body.
3. Problems in involving private sector partners are, however, not unique to Austrian pacts. Many European pacts faced similar problems (see for instance the experiences reported in: EC, 1998).
4. Both the analysis and the strategies have to exist in written form for all pacts and are renewed at regular intervals in a number of provinces . These documents thus serve as an important data source for evaluation.
5. This is available at *www.pakte.at*.
6. In principle three situations could be imagined. First, the strategy could focus exclusively on one policy field (such as, perhaps, exclusively planning active labour market policy measures). Second, many policy fields could be addressed without much integration. Finally, policy fields could be integrated into a coherent strategy.

7. This is actually confirmed by a recent evaluation of the Viennese pact (Leitner *et al.*, 2002) where a respondent answering the question of whether vague goals are a problem is quoted as saying "In my perception this was never a problem. The advantage is it can be implemented much easier, the disadvantage, the outcome is the same as has been already done" (Leitner *et al.,* 2002).

References

AMS WIEN and WAFF (2000), *Territorialer Beschäftigungspakt Wien 2000*, Wien.

BIFFL *et al.* (2000), *Evaluierung des Nationalen Aktionsplanes für Beschäftigung in Österreich*, WIFO, Wien.

CAMPBELL, Mike (2001), "Partnerships in Austria: Enhancing Regional Co-operation in a Decentralised Policy Framework", in OECD (2001).

EUROPEAN COMMISSION (1998), "Territorial Employment Pacts: Good Practice Seminar, Bremen 25-27 January 1998", Seminar Report.

EUROPEAN COMMISSION (1999), *Leitfaden der Europäischen Kommission für die territorialen Beschäftigungspakte 2000-2006*, *www.pakte.at/rahmendokumente.html*.

GEYR, Renate and Erika HESS (2000), "Der arbeitsmarkpolitische Verbund" 12/23, Wien.

HUBER, Peter and Ewald WALTERDKIRCHEN (1999), "Möglichkeiten Regionaler Beschäftigungspolitik in Oberösterreich", Studie des Österreichischen Instituts für Wirtschaftsforschung im Auftrag des Landes Oberösterreich, WIFO, Wien.

KATAJAMÄKI, Hannu (1998), *Beginning of local Partnership in Finland – Evaluation, interpretation and impressions*, Research Institute at the University of Vaasa, Publication No. 76, Vaasa.

LEITNER, Josef (2000), *Machbarkeitsstudie – Schaffung von Arbeitsplätzen innerhalb von Projekten auf kommunaler und (klein) regionaler Ebene*, *www.pakte.at/ rahmendokumente.html*.

LEITNER, Andrea, Angela WROBELSKI, Peter PRENNER, Helmut HOFER, Andreas SCHUH, Helmut MAHRINGER (2002), *Evaluation der Arbeitsmarktpolitischen Maßnahmen des TBP Wien 1999*, Institute for Advanced Studies, Vienna.

TERRITORIALER BESCHÄFTIGUNGSPAKT SALZBURG (2000), Gesamtbericht – *Arbeitsmarktpolitische Initiativen des Landes Salzburg 1997-1999*, Salzburg.

OECD (2001), *Local Partnerships for Better Governance*, Paris.

PAKT FÜR ARBEIT UND WIRTSCHAFT TIROL, PAKT FÜR ARBEIT UND WIRTSCHAFT TIROL – RAHMENVERTRAG, *www.tirol.gv.at/*.

PAKT FÜR ARBEIT UND WIRTSCHAFT TIROL, PAKT FÜR ARBEIT UND WIRTSCHAFT TIROL, Innsbruck.

PAKT FÜR ARBEIT UND WIRTSCHAFT TIROL, ARBEITSMARKTVEREINBARUNG LAND TIROL – ARBEITSMARKTSERVICE TIROL, *www.tirol.gv.at/*.

SACHSE, Irene, Klaus WIRTH and Gerda ZEMAN-STEYRER (2000), "Kommunale Beschäftigungspolitik in Österreich Status Quo – Best Practices", Studie im Auftrag des Bundesministeriums für Arbeit, Gesundheit und Soziales und des ESF, *www.pakte.at/rahmendokumente.html*.

SAIKKONEN, Paavo, "Observations and Comments on the Austrian Peer Review from the Finnish viewpoint", manuscript, Ministry of Labour , Helsinki, Finland.

SCOPPETTA, Anette (1999), "Territoriale Beschäftigungspakte – Vernetzung von Politikbereichen auf regionaler Ebene". In Eichmann, Hochgerner, Nahrada (ed.): *Netzwerke – Kooperation inArbeit, Wirtschaft und Verwaltung*, Wien: Falter Verlag, Soziale Innovation + Neue Soziologie, Band 5, S109-126.

SCOPPETTA, Anette (2000), "Hochkonjunktur für regionale Partnerschaften – Territoriale Beschäftigungspakte im Aufwind". In *LEADER Magazin Österreich*. Nummer 2_2002, Wien, 2002, 29-30.

ZSI (1999), *Leitfaden der Territorialen Beschäftigungspakte in Österreich 2000-2006*, *www.pakte.at/rahmendokumente.html*.

ZSI (1999), *Fragen und Stellungnehmen zu den Territorialen Beschäftigungspakten*, Wien.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 16

A Commentary on the Workshop "Evaluating Territorial Employment Pacts"

by
Hugh Mosley,
Social Science Research Centre,
Berlin, Germany

"Territorial Employment Pacts" (TEPs) are encompassing, area-based networks for the co-ordination and implementation of employment and economic development policies in integrated projects. In both Austria and Italy this approach was first introduced as an EU pilot program and subsequently continued, in somewhat modified form, as an element of national policy. It is important to note that the TEPs are not organisations in a formal sense with, for example, their own budgets and employees but (more or less) voluntary partnerships in which the constituent organisations retain their individual identity. Only a relatively small amount of funding is provided to cover the costs of the development and co-ordination of the TEP programs, which leverages a disproportionately large amount of joint activity by the member organisations.

The TEP networks themselves are extremely varied within and between countries: They differ, for example, in the areas covered [entire provinces (Austria) or local networks (Italy)], in the constellation of public and private actors involved, whether their principal function is co-ordination and planning (Austria) or operating local programs (Italy), and whether their focus is more on employment (Austria) or economic development (Italy).

There are two special rationales for the TEPs that need to be considered in designing an evaluation: 1) it is assumed that they yield synergy effects by creating an encompassing formalised local coalition with shared goals and co-ordinated projects. If no additional program resources are available, this can only be achieved by increasing efficiency and effectiveness through bundling of existing resources, for example, through better co-ordination of local employment strategies. Focusing limited resources on a smaller number of priority targets may also enhance policy outcomes. 2) Another major rationale is to foster the development of social capital in the form of local policy networks, which are assumed to have a favorable long-term impact on local employment and economic development.

The papers presented by Paola Casavola and by Peter Huber on evaluating territorial employment pacts in Italy and Austria, respectively, are both based on ongoing evaluations of national programs. Both emphasize the objective of promoting local networks for implementing economic and employment policies and seek to develop an appropriate evaluation strategy different from that of mainstream program evaluation. As Peter Huber argues: "Evaluating the efficiency of institutions in implementing programs implies a different

counterfactual from that researched in much of the active labor market policy literature." Rather than asking what would have happened if an individual had not participated, "the counterfactual here should answer the question: what would have happened if the institution designing the measure had not existed."

In her study of territorial pacts in Italy Paola Casavola provides a quantitative assessment of the determinants of pact formation based on data from 61 Italian districts. The general assumptions of the Italian program of incentives for the formation of territorial pacts is that local employment and economic development is fostered 1) by promoting dense networks of social and economic co-operation and that 2) there is a positive "agglomeration" effect to combining individual programs within a broader integrated policy project.[1] These interesting assumptions are not, however, the subject of her paper. Instead her evaluation for the Italian Ministry of Economics and Finance focuses on the intermediate goals that are the actual focus of the Italian program: the progress of the integrated projects and the characteristics of the local coalitions. Combining monitoring data on pact performance with social and economic indicators and survey data on the development of the pact coalitions, she investigates a number of key issues for 61 Italian territorial pacts: 1) To what extent is successful pact formation affected by "initial conditions", *i.e.* social and economic conditions in the pact localities? 2) How does the quality of the start-up process affect pact outcomes? In both cases she uses "realization speed" measured in terms of the stepwise take-up of authorized pact funds as the dependent variable. Briefly stated, she concludes that the realization of pacts was not adversely affected by initial conditions, *i.e.* this form of policy governance is also feasible in disadvantaged localities, and that "pacts" in which all relevant actors were involved from the beginning proceeded faster, which suggest that investment in the start-up phase is particularly important for the success of territorial pacts.

Peter Huber's paper reports on the evaluation design and initial results from an evaluation of the Austrian territorial employment pacts as part of a broader evaluation of European Social Fund Objective III programs in Austria being carried out by the Austrian Institute for Economic Research (WIFO). His focus is likewise processes rather than individual or labor market outcomes. His overriding research question is an institutional counterfactual: How would policies of the participating organisations have been different in the absence of the TEP organisation?

The evaluation design focuses on three principal types of questions: 1) Strategy formulation including issues related to the quality of the pact coalition and its success in achieving a strategic integration of the partners; 2) Analysis of implementation with particular attention to conflicts within coalitions and the impact of the TEP on the participating institutional actors as well as the possible long-term "multiplier effects" of "social capital" created

by the TEP experience. He shows the relevance of traditional program evaluation categories (*e.g.* deadweight, displacement) even in a more qualitative, case study of the impact of governance institutions. 3) Finally, under policy-uptake he advocates giving particular attention to the incentives for regional actors to engage in local policy coalitions, which depend on the perceived costs and benefits for the potential coalition partners.

The discussion on the papers at the Vienna conference was largely devoted to informational questions about the Italian and Austrian programs (*e.g.* differences between pacts in Italy and Austria and the types of labor market programs implemented) and technical aspects of the analysis (*e.g.* how success is measured). More general themes related to the distinctive focus of both papers on "governance issues" rather than individual program outcomes and the difficulty of defining and measuring social capital. Both the Austrian and Italian programs are based on assumptions about the importance of encompassing networks for local employment and economic development, which are not directly addressed in either of the papers.

Both papers illustrate the need for a different evaluation strategy when assessing the impact of local governance structures. Standard evaluation methodology in labor market research has focused largely on individual program effects. Based on an explanatory model borrowed from medical research, this type of evaluation has sought to estimate the impact of program "treatments" on individual participants. In this methodological paradigm the employment and earnings of program participants are compared with that of an experimental or quasi-experimental control group of non-participants (the counterfactual) in order to estimate net effects.[2]

This tradition of impact analysis has greatly increased our sophistication in assessing the net effects of labor market programs. However this standard model has two major shortcomings from the point of evaluation of local employment and economic development policies:[3]

1. It tends to assume a simplified model of the policy process in which program "treatments" are highly standardized, underestimating the importance of local program variation. The actual "treatment" that program participants receive may vary considerably for participants throughout the country. In fact labor market programs provide complex services that are, at local implementation, seldom uniform.[4] Moreover, labor market programs in many countries merely provide framework regulations that by design permit a great deal of local variation in program content. This is the case, for example, for many European Social Fund programs, of which the Austrian territorial employment pacts are an example, as well as for major national training and job creation programs in Germany and in other countries with relatively decentralized public employment services.[5] In

sum, program "treatments" are seldom so standardized as the analogy to medical research suggests (Scheirer 1994). Local effects in program evaluation can in principal be examined if the evaluation design provides valid findings at the relevant regional and local levels.[6]

2. More importantly, the focus on individual program effects neglects a range of other evaluation questions that are central to policy-analysis (management and process evaluations, efficiency studies, cost-benefit analyses, etc.) for which other methodologies may be more suitable (see de Koning and Mosley 2001, Palumbo and Calista 1990, Scheirer 1994). For example, structural issues related to the impact of changes in the organisation of public employment service, including financing, governance, and management structures, cannot be addressed in an experimental or quasi-experimental framework. Similarly, evaluation issues related the impact of local governance structures, like the territorial pacts under consideration here, require a different evaluation design.

What's different about evaluating territorial pacts?

1. Although in both countries the territorial employment pacts originate in an EU initiative later incorporated into national policy, the "local" dimension is in the first instance the territorial or area focus of policy, in contrast to individual- or firm-oriented policies. Improvement in individual programs and aggregate outcomes is sought by influencing collective behavior in formulating and implementing policies at the local level. The objective is to transform local implementation structures by promoting broad coalitions in formulating and implementing policies. The central evaluation question here is thus whether and to what extent this local effect is achieved and not, in the first instance, the impact of program "treatments" on individual participants.[7]
2. A second evaluation question is the impact of the programs run by the territorial pacts on local labor market programs. Has the new form of governance (TEPs) had an impact on the level and mix of local (labor market) policy at the local level? Do the actors involved alter their priorities and policy portfolios as a result of their involvement in the TEP coalition? A major rationale for this type of implementation strategy is to increase the efficiency and effectiveness of policy by improving co-ordination among the different agencies and between public and private actors involved in local policy implementation. For example, in most countries different organisations are responsible for labor market policy and local economic development policies. This seems to be a principal function of the Austrian TEP model, which in most cases represent agreements between the regional public employment service offices and the provincial authorities responsible for economic development.

3. Another distinctive evaluation question is the impact of program-induced changes in governance on aggregate policy outcomes in the regions. In order to answer this question a methodology different from the control group method to assess the impact of programs on participants is required. A weakness of this method even in the case of program evaluation is that it does not take into consideration displacement, substitution and other indirect effects at the aggregate level in the labor market. A regional-level aggregate impact analysis can be used to assess effects at the local level measured in terms of indicators of local employment and economic development. It might utilize a time series or cross section framework (or both), or a matched comparison of regional units.[8] An aggregate impact analysis would, for example, also be the appropriate framework for testing the common assumption of these programs that enhancing "social capital" in the form of local policy coalitions engenders better local economic and labor market performance.
4. A final evaluation question is, last but not least, the classical question of program impact on individual participants. Given the heterogeneity of the territorial pacts and the assumption that local implementation matters, a locally-focused evaluation is required, based ideally on data on local labor market outcomes for participants and a control group. In principle evaluation of local programs can be carried out using the same experimental or quasi-experimental methods applied to national programs, although this appears to be relatively infrequent. For example Leitner *et al.* (2002) examine the impact of labor market policy measures of the Vienna Territorial Employment Pact using a quasi-experimental method based on a propensity score matching procedure. Data on program participants and a control group of non-participants were drawn from social insurance records and used to assess the net impact of the Vienna programs on participants' employment and earnings.

Why was an impact evaluation of participation in local employment programs using a control group method possible in this case? 1) The European Social Fund mandates evaluation of its programs, although the actual quality of the evaluations may vary greatly; 2) The economies of scale that make evaluation of smaller programs in some localities too costly were not an obstacle in Vienna; 3) Evaluation expertise was available (Institute for Advance Studies [HIS]), which should be the case in most European countries; 4) The key element appears to have been the ready availability of data from social security records on the work careers of local participants and non-participants, which in most European countries is not the case; 5) noteworthy also is the fact that the final evaluation report of the 1999 local Vienna programs was completed in the Spring of 2002, which is indicative of the trade-off between evaluation quality and the need of program administrators for real-time feedback.

Notes

1. The implicit model of the first assumption is the success story of the Italian industrial districts, which is attributed to their special "social capital" endowment.
2. For an overview see, for example, the paper by Smith in this volume.
3. See also Mosley, H. and E. Sol (2001).
4. Mosley and Degen (1994) discuss the actual variation in services provided even within highly structured UK training programs.
5. Schierer (1994) distinguishes between "aggregate" and "targeted" programs.
6. The variance in net program effects at the local level could serve as a measure of (unobserved) variation in implementation.
7. In fact, the papers by Peter Huber (Austria) and Paola Casavola (Italy) raise the question of what we mean by "local" evaluation, which refers not just to evaluation of local programs but specific local effects.
8. See de Koning (2001) for a general discussion of approaches to aggregate impact analysis.

References

DE KONING, Jaap and Hugh MOSLEY (eds) (2001), *Active Labor Market Policy and Unemployment*, Cheltenham, UK: Edward Elgar.

DE KONING, Jaap (2001), "Models for Aggregate Impact Analysis of Active Labor Market Policy" in de Koning and Mosley 2001.

LEITNER, Andrea, Angela WROBLEWSKI, Peter PRENNERM, Helmut HOFER, Andreas SCHUH, Helmut MAHRINGER (2002), *Evaluation der arbeitsmarktpolitischen Maßnnahmen des TBP Wien 1999*, Vienna: Institut für Höhere Studien.

MOSLEY, Hugh G. and Christel DEGEN (1994), "The Re organisation of Labor Market Policy: Further Training for the Unemployed in the United Kingdom". WZB Discussion Paper FS I 94-205, Wissenschafts-zentrum Berlin für Sozialforschung.

MOSLEY, Hugh and ELS SOL (2001), "Process Evaluation of Active Labor Market Policies and Trends in Implementation Regimes" in de Koning and Mosley 2001.

PALUMBO, D.J. and D.J. CALISTA (1990), "Opening up the Black Box: Implementation and the Policy Process", in Palumbo and Calista (eds.) *Implementation and the Policy Process: Opening up the Black Box*, Greenwood Press, New York, pp 3-16.

PAWSON, Ray and Nick TILLEY (1997), *Realistic Evaluation*, London: Sage.

SCHEIRER, M.A. (1994), "Designing and Using Process Evaluation", in J. Wholey, H. Hatry, K. Newcomer (eds), *Handbook of Practical Program Evaluation*, San Francisco: Jossey-Bass Publishers, pp. 40-68.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 17

A Review of Impact Assessment Methodologies for Microenterprise Development Programmes

by
Gary Woller,
Associate Professor Romney Institute of Public Management,
Brigham Young University Utah, USA

Introduction

Over the last several years, OECD governments have invested millions of dollars in microenterprise development programmes in OECD and lesser developed countries (LDCs). Microenterprise development is based on a couple of underlying premises: 1) self-employment is a key component in creating economic opportunities for low-income persons with otherwise limited employment or earning options, and 2) the primary constraints to productive self-employment among low-income persons are access to capital (loans) and training. For OECD governments, self-employment expands the range of policy options to combat poverty in its many manifestations. For the poor, self-employment expands the range of livelihood and coping options. Self-employment policies appear to offer particular benefits in economies characterised by chronic under or unemployment or by high levels of informal economy participation by the poor.[1]

In theory, programme participants leverage loans and training to start and expand micro and small enterprises thereby generating higher levels of enterprise returns; higher enterprise returns translate into higher household income; and higher household income is in turn invested in improved household socioeconomic well-being. In practice, the specific impacts of microenterprise development programmes are hard to pin down and harder still to measure. Impact assessments require adoption of research methodologies capable of isolating specific impacts out of a complicated web of causal and mediating factors and high decibels of random environmental "noise", as well as attaching specific units of measurement to tangible and intangible impacts that may or may not lend themselves to precise definition or measurement. It is not an easy task.

Nonetheless, microenterprise development competes with other development-employment policies for scarce public funds, and it is reasonable that policymakers should want to know whether microenterprise development is a good investment relative to other policy options. Fortunately, methodologies to assess programme impact do exist. Drawing on principles and experience in the natural and social sciences, impact assessment (IA) methodologies are well-developed and are well-known to researchers. The same methodologies, however, are much less well-known to policymakers, with important implications for policy analysis.

The validity of the findings of any impact assessment is in direct proportion to the validity of the IA methodology used. There exists substantial variation in the methodologies used by IA researchers, but not all IA methodologies are equally valid. As a result, the quality of IA studies runs the gamut. Methodological variation reflects a number of factors, such as researcher skill and inclination, the purpose of the assessment, and resource or environmental constraints. The truth is that most IA researchers work under constraints that require them to make trade-offs between methodological precision and methodological feasibility. The validity of a particular IA study often turns on the validity of the trade-offs made.

Given the methodological issues that inevitably arise during any impact assessment, the ability of policymakers to reach informed decisions regarding the impact of self-employment policies (or any public policy, for that matter) arguably depends to a large degree on their ability to make informed judgments about the validity of the assessment methodologies used and the justifications for methodological tradeoffs made. To further this end, this study examines IA methodologies used in 67 IA studies of 90 microfinance programmes in 31 LDCs and 20 IA studies of 20 microenterprise programmes in 2 OECD countries (19 of them in the United States, see Tables 17.1-17.3).[2]

The rest of the study proceeds as follows. The following section describes the conceptual foundations to IA, followed in the next section by a discussion of the two dominant methodological paradigms and the five methodological approaches that fall within them. The fourth and fifth sections discuss, respectively, the major methodological pitfalls bedevilling IA and other miscellaneous methodological shortcomings common to IA studies. The sixth section offers additional thoughts about judging the methodological rigor of IA studies. The final section offers some policy recommendations.

Before going farther, one clarification is in order. From here on, the term *microfinance* is used to connote microenterprise development in lesser developed countries (LDCs), and the term *microenterprise* to connote the same in OECD countries. The different terminology reflects important distinctions between the two. In LDCs, the microfinance industry is evolving into a full-fledged financial service industry for the poor. Increasingly, microfinance institutions (MFIs) are offering a range of financial services – such as savings, consumption loans, or insurance – in addition to enterprise loans. Historically, the poor in LDCs have not had access to formal financial services of any kind. Notwithstanding, MFIs are discovering a large latent demand for a diversified set of formal financial services among the poor, and they are evolving to meet that demand. Primary among the goals of many microfinance advocates is "financial deepening", or the creation of a system of sustainable financial intermediation for the poor.

Table 17.1. **Summary of reviewed IA assessments**

Reviewed impact assessment[a]	Year	Country	Methodology[b]	Control group[c]	Timeframe[d]	Comment[e]
LDCs						
Ashe and Parrot	2001	Nepal	S, Q	N	CS	
Barnes	2001	Zimbabwe	S, Q	Y	L	
Barnes, Morris, and Gaile	1999	Uganda	S, Q	Y	L	
Bolnick and Nelson	1990	Indonesia	S	Y	CS	
Buckley	1996a	Kenya	S	Y	CS	
Buckley	1996b	Malawi	S, Q	Y	CS	
Buvinic, Berger, and Jarmillo	1989	Ecuador	S	Y	L	
Chen and Snodgrass	2001	India	S, Q	Y	L	
Churchill	1995	So. Africa	S, Q	Y	CS	
Coleman	1999	Thailand	S, Q	Y	L	SB, DS, CL
Coleman	2001b	Thailand	S, Q	Y	L	SB, DS, CL
Copestake, Bhalotra, and Johnson	2001	Zambia	S, Q	Y	CS	
Creevey, Ndour, and Thiam	1995	Guinea	S, Q	Y	CS	
Deardon and Khan	1994	Bangladesh	S	Y	L	
Diagne	1998	Malawi	S	Y	L	F
Dunn and Arbuckle	2001	Peru	S, Q	Y	L	
Goetz and Gupta	1996	Bangladesh	Q	N	CS	
Gupta and Davalos	1993	Jamaica	S	N	CS	
Hashemi, Schuler, and Riley	1996	Bangladesh	S, Q	Y	L	SB, DC, CL
Hulme, Montgomery, and Bhattacharya	1996	Sri Lanka	S	Y	CS	
Karlan and Alexander	2002	Peru	S	Y	CS	SB, DO
Kevane and Wydick	2001	Guatemala	S	Y	CS	
Khandker	1996	Bangladesh	S	Y	CS	SB, DS, CL
Khandker	2001	Bangladesh	S	Y	L	SB, DS, CL
Khandker, Samad, and Khan	1998	Bangladesh	S	Y	CS	SB, DS,C L
Kilby and D'Zmura	1985	Brazil	S	N	CS	BC, CL
		Upper Volta	S	N	CS	BC, CL
		Honduras	S	N	CS	BC, CL
		Dom. Rep.	S	Y	CS	BC, CL
		Peru	S	N	CS	BC, CL
			S	N	CS	BC, CL
Lapar, Graham, Meyer, and Kraybill[f]	1995	Philippines	S	Y	CS	SB
Lapar, Graham, and Meyer[f]	1995	Philippines	S	Y	CS	SB
McKernan	1996	Bangladesh	S	Y	CS	SB, DS
MkNelly and Dunford	1999a	Bolivia	S, Q	Y	L	SB

Table 17.1. **Summary of reviewed IA assessments** (*cont.*)

Reviewed impact assessment[a]	Year	Country	Methodology[b]	Control group[c]	Timeframe[d]	Comment[e]
MkNelly and Dunford	1999a	Bolivia	S, Q	Y	L	SB
MkNelly and Dunford	1999b	Ghana	S, Q	Y	L	SB
MkNelly, Watetip, and Lassen	1996	Thailand	S, Q	Y	CS	
Montgomery, Bhattacharya, and Hulme	1996	Bangladesh	S, Q	Y	CS	
Morduch	1998	Bangladesh	S	Y	CS	SB, DS, CL
Mosely	1996a	India	S	Y	CS	
Mosely	1996b	Indonesia	S	Y	CS	
Mosely	1996c	Bolivia	S	Y	CS	
Mosely	2001	Bolivia	S, Q	Y	CS	
Mosely and Hulme	1998	Bangladesh	S	Y	CS	
		Bolivia	S	Y	CS	
		India	S	Y	CS	
		Indonesia	S	Y	CS	
		Kenya	S	Y	CS	
		Malawi	S	Y	CS	
		Sri Lanka	S	Y	CS	
Mustafa, Ara, Banu, Hossain, Kubir, Moshin, and Yusuf	1995	Bangladesh	S, Q	Y	CS	DS
Neill, Davalos, Kiiru, and Sebstad	1994	Kenya	S	N	CS	
Nelson	1984	Indonesia	S	Y	CS	
Nelson and Bolnick	1986	Indonesia	S	Y	CS	
Oldham, Hadidid, Hussein, Aziz, and Sakr	1994	Egypt	S	N	CS	
Park and Ren	2001	China	S	Y	CS	
Pitt and Khandker	1996	Bangladesh	S	Y	CS	SB, DS
Pitt and Khandker	1998	Bangladesh	S	Y	CS	SB, DS
Pitt, Khandker, Chowdhury, and Millimet	1998	Bangladesh	S	Y	CS	SB, DS
Pitt, Khandker, McKernan, and Latif	1999	Bangladesh	S	Y	CS	SB, DS, CL
Pulley	1989	India	S	Y	L	
Schuler and Hashemi	1994	Bangladesh	S, Q	Y	L	SB, DS, CL
Schuler, Hashemi, and Badal	1998	Bangladesh	Q	Y	L	
Schuler, Hashemi, and Riley	1997	Bangladesh	S, Q	Y	L	SB, DS, CL
Sebstad	1992	So. Africa	Q	N	CS	
Sebstad and Cohen	2002	Bangladesh	S, Q	Y	CS	
		Bolivia	S, Q	Y	CS	
		Philippines	Q	N	CS	
		Uganda	Q	N	CS	

Table 17.1. **Summary of reviewed IA assessments** *(cont.)*

Reviewed impact assessment[a]	Year	Country	Methodology[b]	Control group[c]	Timeframe[d]	Comment[e]
Sebstad and Loza	1993	Egypt	S, Q	N	CS	
Sebstad and Walsh	1991	Kenya	S, Q	N	CS	
Smith	2002	Ecuador	S	Y	L	
		Honduras	S	Y	L	
Steele, Amin, and Naved	2001	Bangladesh	S	Y	L	
Sutoro	1990	Indonesia	S, Q	N	CS	
Todd	1996	Bangladesh	Q	Y	CS	
Vengroff and Creevey	1994	Senegal	S, Q	Y	CS	
Woller and Parsons	2002	Ecuador	S	N	CS	CL
Wydick	1999a	Guatemala	S	Y	CS	
Wydick	1999b	Guatemala	S	Y	CS	
Zaman	2001	Bangladesh	S	Y	CS	SB, F
Zeller, Ahmed, Babu, Broca, Diagne, and Sharma	1996	Cameroon	S	Y	L	CL
		Mali	S	Y	L	CL
		Ghana	S	Y	CS	CL
		Nepal	S	Y	L	
		Pakistan	S	Y	L	CL
		China	S	Y	CS	CL
		Bangladesh	S	Y	L	CL
		Madagascar	S	Y	L	CL
		Malawi	S	Y	L	
OECD Countries						
Ashe and MacIntyre	2002	USA	S, Q	N	CS	
Benus, Wood, and Grover	1994	USA	S	Y	L	SB
Blair and Klein	2001	USA	S, Q	N	L	
Clark and Huston	1993	USA	S, Q	N	L	
Clark, Kays, Zandniapour, Soto, Doyle	1999	USA	S, Q	N	L	
Drury, Walsh, and Strong	1994	USA	S	N	L	
Dumas	2001	USA	Q	N	CS	
Else and Clay-Thompson	1998	USA	Q	N	CS	BC
Himes and Servon	1998	USA	S, Q	N	CS	DO
Kosanovich and Fleck	2002	USA	S	Y	L	BC
Institute for Social and Economic Development	1994	USA	S	Y	L	
Mt. Auburn Associates	1998	USA	S	N	CS	DO
Raheim	1996	USA	S	N	CS	
Raheim and Friedman	1999	USA	S, Q	N	L	
Sekkesaeter	2002	Norway	S, Q	N	CS	
The Roberts Foundation	1995	USA	S	N	CS	

Table 17.1. **Summary of reviewed IA assessments** *(cont.)*

Reviewed impact assessment[a]	Year	Country	Methodology[b]	Control group[c]	Timeframe[d]	Comment[e]
US Department of Health and Human Services	1994a	USA	S	Y	L	
US Department of Health and Human Services	1994b	USA	S	Y	L	BC
US Department of Health and Human Services	1994c	USA	S	Y	L	SB
US Department of Health and Human Services	1994d	USA	S, Q	Y	CS	

a) Several reviewed impact assessments were based in full or in part on the same programme assessment or assessment data. These include Coleman (1999, 2001b); Nelson (1984), Nelson and Bolnick (1986), Bolnick and Nelson (1990); Khandker (1994), Pitt and Khandker (1994, 1998), McKernan (1996), Pitt, Khandker, Chowdhury, Millimet (1998), Khandker, Samad, and Khan (1998), Pitt, Khandker, McKernan, and Latif (1999), and Khandker (2001); Schuler and Hashemi (1994), Hashemi and Riley (1996), Schuler, Hashemi, and Riley (1997), and Schuler, Hashemi, and Badal (1998); Wydick (1999a, 1999b) and Kevane and Wydick (2001).
b) S = Impact survey; Q = Qualitative assessment.
c) Y = Yes; N = No.
d) L = Longitudinal; CS = Cross sectional.
e) SB = Controls for selection bias for unobservable characteristics; DS = Administers survey at different seasons/times during year; CL = Assesses community-level benefits; F = Controls for loan fungibility; BC = Performs benefit-cost analysis.
f) Not a programme assessment, but a study of rural non-farm enterprises.

The same is less true in OECD countries. Relative to LDCs, OECD countries have highly developed financial services markets to which the poor enjoy access (if at times limited). There is, moreover, relatively little discussion among microenterprise advocates of financial deepening; instead, the industry's objectives tend to be defined more narrowly within the context of self-employment as an alternative to formal sector employment. Another important difference between microfinance and microenterprise is that the latter emphasises business training to a much greater degree. Integration of loans with business training is common within the microenterprise industry, but is rarer and is the subject of much dispute within the microfinance industry.

Given the differences that exist between the two, therefore, care is taken here to differentiate between them where relevant. That said, the issues discussed in this study apply more or less equally to impact assessments of both types of programmes. Moreover, the need for sound impact assessments applies equally to both as do the methodological principles for conducting them.

Table 17.2. **Summary of IA studies**

	LDC	OECD	Total
Published impact studies	67	20	87[a]
Impact assessments performed	90	20	110[b]
Methodology			
Survey only	58	10	68
Survey + Qualitative	26	8	34
Qualitative only	6	2	8
Control group			
Yes	73	7	81
No	17[c]	13	28
Timeframe			
Longitudinal	27	11	38
Cross-sectional	63	9	72
Control for selection bias	19	2	21
Surveys at different seasons	14	0	14
Assess community-level impacts	24	0	24
Controls for loan fungibility	2	0	2
Performs benefit-cost analysis	6	2	8
Includes programme dropouts	1	2	3
Number of countries	31	2	33

a) Refers to IA studies published, regardless of the number of separate impact assessments covered in each study.

b) Refers to the number of separate impact assessments found in each published IA study. Impact assessments of two or more programmes in the same country as part of an integrated impact assessment were counted as a single assessment. Impact assessments of programmes in different countries, but published in the same IA study, were counted as separate assessments.

c) Control groups were not relevant to the community economic impact methodology used by Woller and Parsons (2002).

Conceptual foundations to impact assessment

Impact theoretically occurs at four levels: the individual, the enterprise, the household, and the community.[3] In theory, the impact causal chain works something like this: 1) loans and training lead to increased enterprise formation and expansion and to increased investment in working capital and productive assets; 2) increased enterprise formation, expansion, and investment lead to increased enterprise returns; 3) increased enterprise returns lead to increased job creation and increased household income; 4) increased household income leads to higher levels of household consumption, asset accumulation, human resource investment, and physical asset investment. Increased household income and asset accumulation, together with increased access to financial services, in turn expand poor households' *ex ante* and *ex post* coping and livelihood strategies, thereby making them less vulnerable to risk.

Table 17.3. **Countries in which impact assessments performed**

	#
Bangladesh	22
Bolivia	5
Brazil	1
Cameroon	1
China	2
Dominican Republic	1
Ecuador	3
Egypt	2
Ghana	2
Guatemala	3
Guinea	1
Honduras	2
India	4
Indonesia	6
Jamaica	1
Kenya	4
Madagascar	1
Malawi	4
Mali	1
Nepal	2
Norway	1
Pakistan	1
Peru	3
Philippines	3
Senegal	1
South Africa	2
Sri Lanka	2
Thailand	3
Uganda	2
United States	20
Upper Volta	1
Zambia	1
Zimbabwe	1

Moving down to the individual level, access to financial services, control over financial resources, enterprise ownership and operation, increased household income contribution, and group networking and mutual support lead to higher levels of personal and social empowerment, especially among female programme participants. At the community level, benefits created at the other three levels create positive externalities that diffuse through local and surrounding communities. (Table 17.4 lists common indicators used to measure impacts at each of the four levels of impact.)

Table 17.4. **Impact Indicators at the individual, enterprise, household and community levels**

Level of Impact	Indicators
Individual Level	• Intra-household decision making (participation in household decision making on issues such as finances, schooling, healthcare, family planning, etc.) • Control over financial and other resources • Contribution to household income • Contraceptive usage • Self-esteem • Attitudes about self, life and the future • Political and social awareness • Participation in social and political spheres • Spousal abuse
Enterprise Level	• Sales • Profits • Net worth • Asset ownership and acquisition • Jobs created • Product diversification • Business diversification • Business practices adopted
Household Level	• Income – Expenditures – Expenditures on food – Expenditures on specific types of foods (*e.g.* fruits, vegetables, meats, dairy) – Expenditures on medicine and health careExpenditures on children's schooling • Asset ownership and acquisition • Savings • Investment in housing and home improvements • Investment in land • Access to and use of medicines and healthcare • Knowledge and use of simple hygiene practices • Knowledge and use of simple medical interventions/health practices (*e.g.* oral rehydration therapy, breast feeding) • Children's school attendance • Types and frequencies of foods consumed • Incidence and duration of "hungry seasons" • Anthropomorphic measures of children • Response to and impact of economic and other shocks
Community Level	• Children's school attendance • Contraceptive usage • Jobs created/Employment • Income • Expenditures • Net worth • Production • Wages • Prices • Participation in social and political spheres • Contribution to families' support • Poverty

In practice, the impact causal chain is more complex than depicted above. For one thing, a host of mediating factors influence impact. Examples include programme attributes, client characteristics, geography, social structure and power relationships, the physical and economic infrastructure, and the macro economy. Another thing is the reciprocal relationship between cause and effect in which impacts become causes, causes become impacts, impacts become causes again, and so on, such that it becomes increasingly difficult to distinguish between the one and the other. Finally, loan fungibility makes tracing through the exact sequence of cause and effect virtually impossible (see discussion below), thus creating something more akin to a causal web than a causal chain. The end result is that participants experience impacts differently, and no two programmes create the same impacts in the same way. Thus it is perhaps not surprising that the only consistency in IA findings is their inconsistency – a typical impact assessment yields mixed findings, and findings vary considerably from study to study. This inconsistency highlights the inherent dangers in assuming a causal link between programme participation and any specific policy outcome.

Methodological approaches to impact assessment

Conceptual frameworks in hand, IA researchers may choose from among several methodological approaches for conducting impact assessments. Depending on the purpose, approaches fall within one or both of two methodological paradigms: the "proving" paradigm and the "improving" paradigm. Within the proving paradigm, the purpose of IA is to attribute causality of observed outcomes to programme participation. The proving paradigm adopts the language, methodology, and worldview of the physical and social sciences. Its audience is primarily external – donors, policymakers, and academics – for whom methodological rigor and scientific validity are prime virtues.

Within the improving paradigm, the objective of IA is to improve the impact of financial services on programme participants through improving products and policies. The improving paradigm adopts the language, methodologies, and worldview of management. Its audience is primarily internal – board, management, staff, and clients – for whom usefulness, timeliness, and cost are prime considerations.

Within the context of the proving and improving paradigms, five methodological approaches can be identified:

1. The scientific method, which is based in the natural sciences.
2. The humanities tradition, which uses theory and corroboration of evidence to make reasoned judgments.

3. Midrange assessments, which explicitly take into account constraints imposed by "field realities."
4. Participatory Learning and Action (PLA), which facilitates subjective articulation of participants" "reality" to arrive at informed conclusions.[4]
5. Market research, which emphasises the collection and use of market intelligence to inform management decision making.

If we assume a continuum, with the proving paradigm at one extreme and the improving paradigm at the other, the scientific method and market research would lie at or near the respective extremes. The remaining three methods would tend to cluster near the middle of the continuum. The scientific method relies primarily on quantitative evidence and the humanities tradition and PLA primarily on qualitative. Midrange assessments and market research tend to rely more on varying combinations of quantitative and qualitative evidence. In practice, all five approaches may be used for either proving or improving. Moreover, it is not uncommon for researchers to use a combination of approaches, along with combinations of quantitative and qualitative methods, so as to crosscheck data and add greater depth, confidence, or relevance to the findings.

Scientific method

In the classic scientific experiment, study subjects are drawn from the same population, share common characteristics, and are randomly selected into either the treatment group (those receiving the intervention) or the control group (those not receiving the intervention). The intent of the classic experiment is to control for all mediating factors so as to be able to attribute any observed differences in outcomes between treatment and control groups to the intervention. Unfortunately, in the social sciences the conditions for a classic experiment rarely exist, and the ability of the researcher to control for all mediating factors is severely limited, if not impossible.

A step down from the classic experiment is the "quasi-experiment", which attempts to replicate the conditions of the classic experiment to the extent possible within existing constraints and typically using survey-based research instruments. Another way to think of the quasi-experiment is as a "with-without" test, in which the researcher attempts to establish the counterfactual of what would have happened had the treatment group not received the intervention.

Multiple regression is a form of quasi-experiment. It tests for the impact of certain explanatory variables on observed outcomes, while holding other mediating factors constant. Multiple regression is not widely used in impact assessments, however, owing to the large data demands necessary to account for all relevant mediating factors, the technical expertise required to perform

and interpret it, and the tenuousness of the econometric assumptions necessary to validate it.

The most common form of quasi-experiment is the control group study. As the name suggests, control group studies compare outcomes in a treatment group of programme participants to a control group of non-participants. If all goes well, the control group matches the treatment group on key characteristics, and the researcher can reasonably attribute observed differences in outcomes to programme participation. All does not always go well, however. Control group studies are fraught with several potential methodological pitfalls. Stumbling into any of the pitfalls can seriously compromise the study's validity.

Next in order of rigor are impact assessments that aspire to the scientific method, but are more properly characterised as pseudo-scientific, because they violate certain key scientific principles. A common example is the "before-after" assessment, which compares outcomes among programme participants at one point (*e.g.,* after joining the programme) to that of an earlier point (*e.g.,* before joining the programme), but which does not use control groups. The absence of control groups in the before-after assessment makes any attribution of impact to programme participation statistically invalid.

Humanities tradition

The humanities tradition is an inductive approach to learning that encompasses a broad set of tools developed and refined by social scientists in such fields as sociology and anthropology. Borrowing liberally from ethnography, the humanities approach involves the study of a small group of subjects in their own environment. Rather than looking at a small set of variables and a large number of subjects, it attempts to get a detailed understanding of the circumstances of the few subjects being studied. It is descriptive and interpretive – descriptive because detail is so crucial and interpretive because the researcher must determine the significance of what he or she observes without gathering broad, statistical information. Methods include key informant and in-depth participant interviews, case studies, and participant observation (*e.g.,* extended residence in programme communities by field researchers).

The humanities tradition does not attempt to prove impact in any statistical sense, but to offer an interpretation of the relationship between programme participation and outcomes that achieves high levels of plausibility, which in turn permits inference of causality between observed outcomes and programme participation. Plausibility itself depends on factors such as evidence of sound methodology, logical consistency, quality of evidence and reasoning, extent of confirmatory evidence via triangulation and secondary information sources, and the reputation of the researcher.

Relative to the scientific method, the humanities tradition lacks a set of well-defined methodological standards. Whether such standards could be specified to guide researchers *ex ante* and to help policymakers gauge the quality of work *ex post* is an as-of-yet unanswered question. Nonetheless, experience has shown that this approach can yield results of reasonably high reliability and provide a range and depth of insights not always obtainable through scientific methods. It may even at times produce conclusions of greater validity than scientific methods, particularly in the case of "survey based IA work that masquerades as science, but has not collected data with scientific rigor" (Hulme, 2000, p. 87).

Participatory Learning and Action

Participatory Learning and Action (PLA) encompasses a wide variety of methodologies, including Participatory Rural Appraisal (PRA), Rapid Rural Appraisal (RRA), Participatory Learning Methods (PALM), Participatory Action Research (PAR), or Farming Systems Research (FSR). PLA assumes the existence of multiple subjective realities. It requires first that researchers answer the question "whose reality counts"? (Chambers 1997). The job of the researcher is next to elucidate the shared reality of the target group through the process of "knowledge creation", defined as the full participation of the target group in problem identification, analysis, and action planning. PLA methods include activities such as visualisation, seasonal calendars, historical timelines, Venn diagrams, transect walks, focus group discussions, relative preference ranking, and semi-structured interviews.

PLA offers the most radical and serious challenge to the scientific method as applied to international development. PLA advocates pull no punches in listing what they see as the flaws of the scientific method (Hulme, 2000):

- It ignores the complexity, diversity, and contingency of poor households' livelihood and coping strategies.
- It conceives causality as a simple, unidirectional chain and not as the complex web that it is.
- It measures the trivial or pretends to measure what cannot be measured.
- It is extractive and exploitative.
- It reinforces the status quo through empowerment of technocrats, experts, professionals, policy-makers and elites.
- It does not lead to purposive action by or on behalf of poor groups.

If true, these are damning criticisms. Whether these allegations are true is perhaps a matter of one's own subjective reality, although they probably do contain a good deal of truth. At the very least, they call for a certain level of

humility and critical self-reflection among practitioners of the scientific method, two traits that are not always abundantly evident.

PLA has its own set of weaknesses. Among them are its inherent subjectivity, its lack of standardisation (which makes comparisons difficult), its pluralism (which produces conflicting perspectives about impact), its naïve assumption that participation equals representation given local power relations, its assertion that participation is tantamount to empowerment, and its lack of transparency (which makes *ex-post* evaluations of methodological rigor difficult).

A final issue is attribution. Much like the humanities tradition, PLA neither can nor does claim causality on purely scientific grounds. It relies on triangulation of evidence, depth of knowledge, quality of methods and information, and skill of the researcher to establish a plausible case for causality. Such limitations, however, do not deter its advocates from arguing that well-conducted participatory studies can produce more reliable results than conventional surveys (see, for example, Chambers, 1997, pp. 141-146.)

As in the humanities tradition, the quality and reliability of PLA studies vary widely, depending on factors such as the skill of facilitators, the motivation of the target groups, the applicability of tools to situations, or the degree of participation.[5] In practice, PLA advocates have been remarkably pragmatic. They are hesitant to prescribe specific best practices, preferring to rely on practitioners' best judgments to adapt the methodology and tools to the circumstances.

Midrange assessments

Midrange impact assessments stem from microfinance practitioners' widespread perception that the scientific method is disconnected from both the realities of the field and the needs of management. Practitioners have long complained that survey-based impact assessments are costly, lengthy, burdensome, and not timely; they require technical expertise and resources beyond institutional capabilities; and they are targeted to external audiences, with little attention to managerial usefulness. Nor do funding agencies or external evaluators invest time or money building institutional IA capacity. Throw into the mix the pressure from donors to achieve institutional self-sufficiency, and IA comes to be seen primarily as yet another line item on the expense report. Lacking the resources, technical skills, and material incentive to conduct IA, most programmes do not do it, the result being that most have little to no idea what their programme impact is.

Midrange assessments are the product of a practitioner-led effort, coordinated by the AIMS project (see footnote 2), to correct the deficiencies of the scientific method, bridge the proving and improving paradigms, and build

institutional IA capacity. The end result of this effort was the creation of a set of "practitioner-friendly" IA tools designed to account for field realities, produce managerially useful results, and be implemented by programme staff. The SEEP/AIMS tools, as they have come to be called,[6] incorporate methodologies from each of the other four IA methodological approaches. They consist of two quantitative and three qualitative tools: the impact survey, the client exit survey, client satisfaction focus groups, in-depth empowerment interviews, and savings and loan use over time interviews.[7] Midrange assessments do not aspire to proof of impact but instead aim to establish "plausible association" between observed outcomes and programme participation.

The SEEP/AIMS tools have enjoyed respectable legitimacy among practitioner organisations since their introduction. To date, dozens of MFIs have received formal, intensive training in use of the tools, and several have in turn successfully implemented them in the field. Midrange assessments, however, are not limited to the SEEP/AIMS tool, nor are all practitioners satisfied with them. Work continues to refine the SEEP/AIMS tools, adapt them to local contexts, or develop yet even more practitioner-friendly tools (*e.g.*,"AIMS-Lite").

For the most part, midrange assessments adhere to the standards of methodological approaches from which they borrow. The major concessions they make are to recommend the use of programme staff to conduct research and, related to the impact survey, to conduct cross-sectional (as opposed to longitudinal) assessments using so-called "pipeline" clients – clients recruited through normal programme operations but who have not yet received loans – as the control group. (See the Appendix for a list of recommendations for conducting midrange assessments.)

Market research

Integral to the improving paradigm is that impact assessment and market research are inextricably intertwined: timely knowledge about impact tells programme management how effective its products and policies are; market knowledge in turn allows programme management to design products and policies that improve impact. Market research (defined as the collection, analysis, and use of market intelligence) therefore plays a central role within the improving paradigm. The emergence of market research on the microfinance agenda is due to several market trends: 1) competition in the industry is increasing and is expected to increase yet more; 2) microfinance consumers are becoming more knowledgeable, discerning, and assertive; 3) clients are deserting microfinance programmes at often alarming rates, and, consequently, 4) MFIs are adopting more commercial strategies and practices.

Market research uses a variety of quantitative and qualitative tools, including surveys, focus group discussions, in-depth interviews, and participatory assessments.[8] Market research includes the occasional assessment activity and proceeds on up to the integration of client and market information into programmes' operational and management information systems. Market research makes little pretence to scientific validity. Lack of attribution is a particular weakness, particularly given the logistical difficulties of integrating a non-client control group into routine data-gathering systems. From management's perspective, however, the loss in statistical certainty is more than made up for by gains in speed, cost, and usefulness. Lagging behind tool development in market research is the development and establishment of a set of methodological standards and guidelines to implementation of market research tools in the field (for more on the relationship between IA and market research, see Cohen 1999 and Copestake 2000.)

Among the five methodological approaches reviewed above, the scientific method and midrange assessments have the most clearly articulated standards to guide methodology *ex ante* and to evaluate the quality of methodology *ex post* (given midrange assessments' aspiration to scientific plausibility, they are subject to many of the same methodological standards as the scientific method). Moreover, survey-based scientific methods dominate IA studies. In most IA studies, qualitative methods (whether grounded in the humanities tradition or PLA) are used as a supplement to survey-based scientific research. Of the reviewed IA assessments, 68 used surveys as the sole research instrument and 34 used surveys as the primary research instrument, complemented by qualitative methods. Only 8 reviewed assessments relied solely on qualitative methods, including Goetz and Gupta (1996), Todd (1996), Schuler *et al.* (1997), Schuler *et al.* (1998), Dumas (2001), and the programme assessments in the Philippines and Uganda summarised in Cohen and Sebstad (2000). In light of the dominance of scientific IA and its relatively clear methodological standards, the rest of this study focuses on methodological issues relevant to scientific IA.

Methodological pitfalls of scientific IA

As mentioned above, ideal conditions to conduct scientific IA rarely exist. IA researchers must therefore often settle for second best, or worse. In the face of ever-present field constraints, IA researchers have to ask themselves to what degree are they willing to compromise accepted methodological principles to accommodate these constraints. In other words, "What cost in scientific precision are IA researchers willing to accept in exchange for a corresponding gain in implementation feasibility?"

In answering the question about acceptable tradeoffs, it is useful to know what the major pitfalls are to conducting scientific IA. This section reviews these pitfalls. In order of presentation, they are 1) construction of valid control groups, 2) selection bias stemming from observable individual characteristics, unobservable individual characteristics, failure to account for programme dropouts, and non-random programme placement; 3) control group contamination; 4) recall bias; 5) loan fungibility; and 6) IA timeframe (longitudinal vs. cross-sectional studies).

Valid control groups

Any scientific impact assessment claiming to infer causality or plausible association between outcomes and programme participation requires comparison to a valid control group of non-clients. Constructing a valid control group, however, can be difficult. There are many reasons why construction of a valid control group may not be feasible, primary among them being binding resource or environmental constraints (common issues for practitioner-led impact assessments) and, as explained below, challenges for control group construction and tracking posed by longitudinal (time-series) impact assessments.

Whether because of binding constraints or other reasons, several impact assessments reviewed for this study did not use control groups. The absence of control groups was particularly conspicuous among microenterprise programme assessments; only 7 of 20 microenterprise programme assessments used a control group compared to 73 of 90 of microfinance programme assessments.

The inability to infer causality does not mean that impact assessments lacking valid control groups have no value, but that their value lies elsewhere. They can, for example, be valuable tools for monitoring client progress, assessing relative outcomes among different market segments, measuring outcomes against programme objectives, or calculating per unit costs for specific programme outcomes or outputs, all of which are useful information for programme management, donors, and policymakers. They are not useful, however, for determining whether programme participation is causally linked to desired policy outcomes, such as poverty reduction or job creation.

Basically, control group selection requires identification of a population of persons not participating in a credit or training programme and sharing similar characteristics as the treatment group and then randomly selecting from among them. Not everyone who belongs to the population of non-clients, however, is a legitimate control group candidate. Construction of a valid control group requires that the control group match the treatment group on key observable and unobservable characteristics. Failure to do so creates so-called

"selection bias", which is the major methodological pitfall bedevilling control group studies. Depending on its seriousness, selection bias renders attribution of observed impacts anywhere from problematic to wholly invalid. Twenty of 90 LDC programme assessments attempted to control for selection bias, compared to 2 of 20 OECD programme assessments.

Selection bias

Selection bias stems from four principals sources: 1) failure to match treatment and control groups on observable individual characteristics, 2) failure to match treatment and control groups on unobservable individual characteristics, 3) failure to account for programme dropouts and 4) non-random programme placement. These are considered in turn below.

Failure to match on observable individual characteristics

Outcome differences between treatment and control groups may be as much, if not more, a function of differences in observable characteristics, such as gender, age, education, self-employment status, enterprise type, or geographic location, than programme participation. Construction of a valid control group thus requires that control group members share similar observable characteristics as treatment group members. For example, if the profile of the treatment group is 85 per cent female, 25-50 in age, 0-5 years of formal education, self-employed, rural, and more or less evenly distributed among manufacturing, retail, and services, a valid control group will match these observed characteristics as closely as possible.

Matching treatment and control groups on observable characteristics can be challenging, but it is by no means insurmountable. A good researcher should be able to avoid this pitfall with comparative ease. To ensure that the control group closely matches the treatment group, sample stratification – random sampling within specifically selected groups among the target population – may be necessary. Whether the treatment and control groups reasonably match on observable characteristics can easily be determined by comparing group means on key observable characteristics and testing whether the differences in means are statistically significant.

Failure to match on unobservable individual characteristics

A yet more bedevilling source of selection bias is failure to match treatment and control groups on unobservable individual characteristics that might also have an impact on outcomes. One might for example ask, "Why does one person join a microfinance programme and another not?" Or "Why are some people early joiners and other people late joiners?" The answers to these questions are probably multifaceted, but a reasonable hypothesis is that,

on balance, joiners and early joiners possess some unseen characteristics that non-joiners or late joiners do not, whether those be entrepreneurial drive, willingness to assume risk, a supportive home environment, or simple determination to improve one's life. The answer may also reflect expected net benefits of programme participation. Joiners and early joiners arguably have higher *ex ante* expected net benefits of participation than non-joiners or late joiners. They will, as a result, also tend to enjoy higher *ex post* net benefits. The point is that unobservable individual characteristics are quite possibly key in determining the impacts of programme participation.

The best method to control for selection bias stemming from unobservable characteristics is through random assignment of study participants into treatment and control groups. Random assignment among IA studies, however, is rare, suggesting that IA studies routinely overstate programme impact. The few IA studies using random assignment methods include Benus *et al*. (1994), US Department of Health and Human Services (1994c), MkNelly and Dunford (1999a, 1999b), and Coleman (1999b, 2001).

A good example of random assignment is Benus *et al*.'s (1994) assessment of self-employment demonstrations in Washington state and Massachusetts. Researchers invited unemployment claimants interested in self-employment to attend an information session that explained basic information about the risks and rewards of self-employment and the key features of the demonstration. At the conclusion of the session, claimants still interested were given an application for the programme. Those who completed the application on time and met eligibility requirements were then randomly assigned either to a treatment group eligible to receive business development services and financial assistance or to a control group that was not.

A simpler method to control for selection bias is to use pipeline clients as the control group, as recommended by the SEEP/AIMS tools. Because pipeline clients have self-selected themselves into the programme, they are presumed to share the same unobserved characteristics as existing clients. Pipeline clients have also been used as a control group in several published academic IA studies (Buckley 1996a, 1996b; Montgomery *et al*. 1996; Mosely 1996a, 1996b, 1996c, 2001; Mosely and Hulme 1998; and Copestake *et al*. 2001).

The use of pipeline clients, however, suffers from some important methodological weaknesses. First, it is most appropriate for cross-sectional studies. A longitudinal study requires that the control group, or at least part of it, not receive the treatment during the entire period of the study. It is probably not operationally feasible in most cases to withhold loans for the duration of a longitudinal study from new clients recruited through day-to-day programme operations. The exception is pipeline clients recruited specifically as part of a controlled, longitudinal IA study.

Second, pipeline clients do not account for why some people join early and some join late. Arguably, someone who joins the programme at or soon after its inception is different from someone who joins two years later. This is less of a problem for those pipeline clients who did not have the option to join earlier, who did not join earlier for reasons unrelated to unobservable determinants of success, or who are drawn from communities where the programme is not currently operating.[9] The point is that in most cases, we do not know why late joiners have waited to join, and almost certainly the explanation frequently involves one or more unobservable traits. In summary, pipeline clients are a far from ideal control group; most will concede that a control group of non-clients is strongly preferable. Advocates of this approach, however, argue that it is a practical solution to situations in which significant or binding constraints exist.

A minimal approach to account for selection bias requires that control group members be drawn from the population of microentrepreneurs who are candidates to join the programme. Better yet is that control group members are drawn from the population of microentrepreneurs who satisfy specific programme eligibility requirements. The weakness of these approaches of course is that they assume incorrectly that all microentrepreneurs or those microentrepreneurs eligible to join the programme would self-select into the programme. Still, the probability that microentrepreneurs eligible to join the programme would self-select into the programme is greater than non-eligible microentrepreneurs. Thus while this approach will not eliminate selection bias, it will at least tend to reduce its prevalence.[10]

Given the large number of self-employed poor toiling in the informal sector in LDCs, there exists a large pool of microentrepreneurs who are legitimate candidates or who satisfy eligibility requirements to join microfinance programmes. Thus there is little justifiable reason for IA researchers in LDCs not to select the control group from this pool. In contrast, OECD countries have much lower incidence of self-employment or informal sector activity among the poor, which makes control group selection from among the pool of self-employed poor that much more difficult. In fact, none of the microenterprise assessments reviewed here drew their control group from this pool. Instead control group members were drawn from Temporary Aid for Needy Families (TANF) recipients (Raheim and Salome 1999), welfare recipients (Raheim and Friedman 1999), unemployment benefit recipients (Kosanovich and Fleck 2002), food stamp recipients (US Department of Health and Human Services 1994b), and Aid for Dependent Children (ADFC) recipients (Institute for Social and Economic Development 1994, US Department of Health and Human Services 1994c). Comparison of microenterprise programme participants to a random sample of government aid recipients almost certainly produces selection bias and overstatement of programme impact.

Failure to account for programme dropouts

Of the reviewed impact assessments, only Karlan and Alexander (2002), Himes and Servon (1998), and Mt. Auburn Associates (1998) included programme dropouts among survey respondents (neither Himes and Servon nor Mt. Auburn Associates, however, used a control group of non-clients.) Omitting programme dropouts from the treatment group introduces two potentially serious sources of selection bias, what Karlan (2001) refers to as *incomplete sample bias* and *attrition bias*. Incomplete sample bias arises because dropouts presumably have fared differently, and quite possibly worse, than those who remain. In contrast, the control group (whether non-clients or pipeline clients) includes some who will succeed and some who will fail. Thus IA studies that omit dropouts from the treatment group compare the programme's successes to a control group of both successes and failures. Consequently, incomplete sample bias produces systematic overstatement of programme impacts.

Attrition bias arises if dropouts are systematically different from those who remain, regardless of impact. If, for example, richer members tended to drop out more than poorer members, then the treatment group would include a higher percentage of poorer members than the control group, and *vice versa* if poorer members tended to drop out more than richer members. The result is systematic understatement of programme impact in the first case and systematic overstatement of programme impact in the second case.

To test the effect of incomplete sample and attrition bias on impact assessment findings, Karlan and Alexander (2002) compared a treatment group from a Peruvian MFI minus dropouts to a control group of the MFI's pipeline clients (as per SEEP/AIMS recommendations) and found statistically significant evidence of positive impacts. After adding dropouts back into the treatment group and comparing the two groups again, they found that the positive impacts disappeared.

Non-random programme placement

Programme placement is not random. Programme managers presumably base programme placement on a variety of strategic criteria, for example, consistency with institutional mission (*e.g.*, high density of very poor), logistical feasibility (*e.g.*, within reasonable distance of programme headquarters), financial attractiveness (*e.g.*, high density of self-employed), or likelihood of successful implementation (*e.g.*, relatively well-developed infrastructure or better access to markets). It is reasonable, moreover, to assume that MFIs will begin in and expand first to those locations that best satisfy the strategic criteria. Such locations arguably share a set of observable and unobservable characteristics that would not be present to the same

degree in a random sample of other locations. These characteristics in turn are potentially significant in explaining relative outcomes from programme participation. Thus, controlling for selection bias also requires selection of a control group from communities that share similar observable and unobservable characteristics as the treatment group community.

The best way to control for non-random placement bias is to randomise programme placement. This approach may appear administratively impractical, but it need not be, particularly if designed to exploit natural limits to programme expansion. Consider, for example, an MFI that plans to expand to *x* number of locations over the next two years. All *x* locations satisfy the programme's placement criteria. Due to resource constraints, however, the programme cannot expand to all *x* locations at once, so its plans call for it to expand to *y* locations this year and *x-y* locations next year. Since the MFI is largely indifferent to the order of expansion, randomizing the selection process both satisfies the programme's expansion criteria and controls for non-random placement bias.

The only examples of randomised programme placement among the IA studies reviewed here are MkNelly and Dunford (1999a, 1999b) in assessments of microfinance programmes in Bolivia and Ghana. Both assessments followed the same methodology. Programme management selected communities to which it would expand over the next two years. Programme staff next visited each community to recruit participants into the programme. In each instance, the programme staff made clear that the community might be assigned to the control group that would not received the programme for two years. Programme staff then collected baseline data from all eligible persons who elected to join the programme. Following baseline data collection, researchers stratified the study communities according to key community characteristics such as size, access to the main road, distance from a market, and access to water. Finally, researchers randomly assigned communities to control and treatment communities in a way that minimised the differences between key community characteristics.

Coleman (1999, 2001b) used a somewhat different approach in his assessment of two microfinance programmes in Thailand. The treatment community included eight villages with access to the programmes for two to four years. The six control communities were pre-selected to receive the programme one year after they were identified. Villagers in the control communities self-selected whether to participate in the programme. To account for the possibility that the order in which the fourteen villages received the programme was not random, Coleman collected a third sample of non-participants in each of the villages. Coleman found that naïve estimates that did not account for self-selection or non-random programme placement significantly overestimated programme impact on several outcome variables,

including overall wealth, land holdings, non-land farm assets, savings, and household income. He concluded that unobservable characteristics, not participation in microfinance programmes, were the most significant determinants of small business income.

Econometric techniques may also be used to control for selection bias. An example is the approach used originally by Pitt and Khandker (1994) to assess the impact of three Bangladesh programmes (and subsequently used by McKernan 1996; Pitt *et al.* 1998; Pitt and Khander 1998; Khandker *et al.* 1998; Pitt *et al.* 1999; Morduch 1998; and Khandker 2001). Their approach exploited programme rules that excluded households with more than a fixed amount of assets from programme participation. In effect, they compared outcomes between eligible and ineligible households in programme villages and outcomes between eligible and ineligible households in non-programme villages and then compared the two differences to each other. They attributed any difference between these two differences to programme participation. Like Coleman, their study showed that naïve estimates that fail to account for selection bias significantly overestimate impact (other IA studies using econometric techniques to control for selection bias include Lapar, Graham, and Meyer 1995; Lapar *et al.* 1995; Zeller *et al.* 1996; and Zaman 2001.) The downside of econometric approaches like Pitt and Khandker's is the sophistication of their econometric methods, which are accessible only to a relatively small group of equally sophisticated methodologists, but not accessible to most policymakers and most certainly not to the average practitioner.[11]

Contamination bias

Contamination bias occurs when the control group becomes contaminated by contact with the treatment group. Contamination can occur in several ways; for example, if control group members are acquainted with treatment group members, members of the two groups share acquaintances in common, the programme initiates contact with control group members or *vice versa* , or knowledge of the programme spreads through formal or informal social networks. Contamination also occurs when programme participation creates positive or negative externalities that influence the behaviour or outcomes of non-participants.

Once control group members become contaminated, the researcher can never be certain whether and how the control group's behaviour and other outcomes have been influenced as a result. To the extent contamination produces better or worse outcomes among treatment group members, it will create systematic understatement or overstatement of actual programme impacts. Contamination bias may be dealt with easily enough by locating the control group away from the treatment group. The farther away the two, the

less the chance of contamination there is (although this has to be weighed against the increase in cost and logistical hassle).

Recall bias

Survey research requires that respondents recall information about their socioeconomic conditions, behaviours, attitudes, and social relationships. Recall responses may be biased, for several reasons. Survey respondents display natural and reasonable tendencies to 1) want to please the interviewer, cast themselves in a good light, or avoid revealing embarrassing information, 2) suspect researchers' motives, 3) game the process, 4) make wild guesses, or 5) give any answer to avoid prolonging unwanted intrusions on their time. Generally, the more personal or intrusive the questions (*e.g.*, sexual practices) or the more difficult to estimate (*e.g.*, household income), the more researchers can expect inaccuracies and biases to creep into the responses.

Using programme staff to interview clients, as recommended by the SEEP/AIMS tools, substantially raises the risk of response bias, and using staff to interview their own clients practically guarantees it. This is not to suggest that programme staff should not be used. In fact, using programme staff to conduct research is often a practical and necessary concession to programme constraints. Moreover, integrating IA or market research into programme systems or weaning one's self from external evaluators may very well require use of field staff to collect impact data. Using field staff to interview their own clients, however, should be done only as a last resort. Programmes using field staff to conduct impact research need to be fully apprised of the risks it poses and be prepared to do what it can to mitigate those risks. The best approach to mitigate these risks is to mix in heavy doses of training, monitoring, and cross-checking of data.

Even where respondents are inclined to give good faith responses, they may or may not be able to recall information with reasonable accuracy. The longer the time period elapsed, the more difficult to estimate accurately. Seasonality can also play havoc in that responses differ depending on the time of year. Strategies to mitigate this source of bias are to use shorter time periods and to survey at different times of the year to reflect key seasons or crop cycles (see, for example, Coleman 1999, 2001b; Pitt *et al.*, 1998; and Mustafa *et al.* 1995).

Loan fungibility

Loans received by programme participants are typically intermingled with other sources of household income, to be spent according to the household's livelihood needs and spending priorities. In other words, loans are not necessarily earmarked for investment in participants' enterprises. In

similar manner, funds for loan repayment do not necessarily come from enterprise cash flows but often come from other sources of household income. Loan fungibility greatly complicates the ability of IA to make a direct link between receipt of a loan and changes in household income, consumption, asset accumulation, individual empowerment, etc. The current state of impact assessment methodology includes no established procedure to account for loan fungibility.

Loan fungibility becomes less a problem, however, if the focus of IA is the household economic portfolio (Chen and Dunn 1996). The concept of household economic portfolio explicitly recognises the fungibility of money as a vital component of poor households' livelihood and coping strategies. It is less interested in the route of the causal chain of impact per se than in the how the impacts ultimately manifest themselves at the different levels of analysis.[12]

Another problem caused by loan fungibility is to make even honest responses to survey questions misleading. To illustrate this point, Coleman (2001a) gives the example of a programme participant who uses her programme loans to pay her children's school fees in place of selling assets to pay the fees, as is her normal practice. In this case, the true incremental benefit of programme participation to the woman is the preservation of assets, which is not measured, and not payment of school fees, which is measured.

Of the studies reviewed here, only Diagne (1998) and Zaman (2001) explicitly deal with loan fungibility. Diagne's approach was to circumvent the problem by making the relevant treatment access to credit rather than receipt of credit based on the reasoning that changes in outcomes because of access to credit were easier to isolate and identify than changes in outcomes because of receipt of credit. Zaman in contrast used an econometric approach based on a household economic portfolio model. His model assumes that money borrowed is spent as needed by utility maximizing households and that by controlling for other factors through application of econometric procedures it is possible to attribute specific outcomes to receipt of the loan. Another possible approach to account for loan fungibility is to collect information on use of loan funds and sources of repayment through additional survey questions or qualitative research.

Longitudinal vs. cross-sectional assessment

Presumably, impacts occur over time, and longitudinal assessments that track outcomes over different points in time shed more light on this process and how it unfolds than do cross-sectional assessments performed at a single point in time. If a longitudinal study is not possible, researchers can proxy a time-series with a cross-sectional assessment by purposefully selecting participants with specific years of experience in the programme and

comparing them either to a control group of non-clients or to pipeline clients. The SEEP/AIMS tools, for example, recommend that researchers select participants with one and two years of programme experience as treatment group members under the assumption that these represent useful points in time where significant impacts might be observed.[13]

The virtues of cross-sectional impact assessments are relatively low cost and low data collection demands. In longitudinal assessments, panel attrition (treatment and control group members dropping out of the study) can also be a problem.[14] Typical ways to deal with panel attrition are tracking down panel dropouts and/or sampling a larger number than otherwise needed, both of which add yet more costs relative to cross-sectional studies.

Despite preference for longitudinal assessments, several cross-sectional impact studies have been published by well-respected researchers or in well-respected academic journals, and cross-sectional assessments make up the bulk of microfinance programme assessments; 63 microfinance assessments were cross-sectional compared to only 27 longitudinal assessments. In contrast, 11 of the reviewed microenterprise assessments were longitudinal and 9 were cross-sectional. Where longitudinal assessments are not feasible, case cross-sectional assessments are an acceptable second best alternative. Best practice holds, however, that where longitudinal assessments are feasible, they should be done.

Other methodological shortcomings

The previous section reviewed the methodological pitfalls common to survey-based impact assessments. Understanding these pitfalls, their causes, and their cures is essential for policymakers to evaluate the validity of IA studies so as to determine their usefulness in informing public policy. In addition to these pitfalls, impact assessments suffer from a variety of other methodological shortcomings that also affect their usefulness for public-policy decision making. These methodological shortcomings have less to do with issues of scientific precision than with providing policymakers with a thorough assessment of the benefits and costs of programme participation.

Ideally, scarce public funds are allocated to those social programmes that yield the highest net social welfare, where net social welfare is defined as the present value of programme benefits minus the present value of programme costs. Estimating net social welfare in turn implies the following: 1) identification of all relevant programme benefits and costs, 2) estimation of relevant programme benefits and costs, 3) conversion of relevant benefits and costs into standardised units so as to permit comparisons within and across programmes, and 4) weighting of relevant programme benefits and costs to reflect social values and priorities. In contrast, most IA studies give only a very

limited picture of programme benefits and virtually no information on programme costs; they do not attempt to convert benefits into standardised units, and they treat all benefits and all costs as equal. Each of these shortcomings, and its implications, is discussed briefly below.

Omission of programme benefits

All impact assessments make choices about which benefits to examine. Most examine some combination of enterprise and household level benefits, many also examine individual level benefits, but nearly all omit community-level benefits, even though evidence suggests that they can be significant in the aggregate.[15] The choice of benefits examined may be based on a number of criteria, for example, donor or programme priorities, industry convention, or personal interest. But whatever the criteria, the end result is that important benefits are inevitably omitted from programme assessments

Omission of important programme benefits gives policymakers an incomplete picture of programme impact. Absent this, the only way for policymakers to form a complete picture of programme impact is to cobble together findings from assorted programme assessments that examine different types of benefits at different levels of analysis. While this approach can be helpful, it can at best give only a very broad picture of impact, and its usefulness in assessing the impact of a particular programme is limited, given the significant contextual disparity in which programmes operate.

Omission of programme costs

Most impact assessments do not mention, let alone estimate, programme costs. Relevant programme costs include the present value of administrative costs and the monetary and opportunity cost of donated/invested funds, soft liabilities, grants-in-kind, price and non-price transaction costs borne by programme participants, and displacement costs (benefit incurred by programme participants at the expense of non-participants). Calculating direct programme costs and price costs to participants is relatively straightforward, but estimating grants-in-kind, participant non-price transaction costs, and displacement costs will tax even the most conscientious researcher, which perhaps explains why so few have done it.

Take non-price transaction costs as an example. Group lending programmes typically require participants to form groups, participate in weekly or biweekly group meetings, and monitor and enforce group loan performance. Such non-price transaction costs are hard to measure, even though they impose significant burdens on programme participants. Harder to measure yet are displacement costs, which can occur, for example, when programmes draw large numbers of the self-employed into sectors attractive

to low-skilled microentrepreneurs and characterised by low barriers to entry, high competition, and low profits. Displacement costs can be significant, reaching as high in some cases as one-half the net benefits accruing to programme participants (Bendick and Egan, 1987). Of the reviewed impact assessments, only Kosanovitch and Fleck (2002) consider displacement costs.

To the extent programme costs are available or can reasonably be estimated, they should arguably be included in impact assessments, if only to provide some baseline for comparison, such as cost per outcome. If costs cannot reasonably be estimated, researchers might at a minimum be expected to identify potential costs, perhaps give some estimate of their order of magnitude, and explain how they might affect the analysis were they to be included.

Kilby and D'Zmura (1985) is the sole microfinance programme assessment reviewed to conduct a benefit-cost analysis. Kilby and D'Zmura assess direct and indirect benefits primarily as measured by value added to assisted enterprises and value added to enterprises outside the project. The former consist of wages, rent, interest, and profit adjusted for the opportunity cost of labor,[16] and the latter of purchases of factors of production made by assisted enterprises and purchases of consumer goods made with direct factor income earned by assisted enterprises. Other benefits considered include training, price reduction, diversion benefits (the benefits derived from diverting a microenterprise loan to another purpose), and weighted wages for the very poor. Costs assessed include all administrative expenditures, bad debt, and capital erosion (the effective interest rate below the rate of inflation).

Several reviewed microenterprise programme assessments conducted at least some form of benefit-cost analysis. Else and Clay-Thompson (1998) calculated the cost per unit of output for clients served, jobs and businesses created, businesses assisted, and loans made. Similarly, the US Department of Health and Human Services (1994) compared the increase in food stamp earnings of demonstration participants to programme administrative costs and the increases in public assistance payments to demonstration participants.

The most complete benefit-cost analysis performed among reviewed microenterprise assessments is Kosanovich and Fleck's (2002) assessment of Self-Employment Assistance Programmes in Maine, New Jersey, and New York. The authors evaluated benefits and costs to programme participants, state governments, and non-participants. Participant benefits assessed included the income gain from self-employment or wage/salary employment, professional development, work satisfaction, and community economic development. Government benefits assessed included increased tax revenue and the reduction in welfare transfers. Participant costs assessed were financial costs of programme administration, training, counselling and the opportunity costs borne by participants who forgo work search and possible

reemployment opportunities while pursuing self employment. Government costs assessed were unemployment insurance payments, programme administration, training, and counselling. Finally, displacement impacts assessed included changes in unemployment insurance benefits and tax payments benefiting participants but paid for by non-participants and the reduction in long-term welfare or unemployment insurance payments benefiting non-participants but at the cost of participants.

Lack of common standards for comparison

Comparing benefits and costs within and across programmes (whether to another microfinance or microenterprise programme or to some other type of development or employment programme) requires that benefits and costs be converted into standardised units. Typically, this entails conversion into monetary units. Without some common standard, policymakers have no way other than their own subjective guesstimates to compare, say, the value of a job created to an increase in a participant's self-esteem in one programme or the value of jobs created by the same programme to the value of increased participant access to health care in another programme. Standardisation of benefits and costs has the added advantage of forcing researchers to be explicit about their assumptions and judgments, thereby both improving analysis and facilitating *ex post* assessments of methodological rigor.

Failure to weigh benefits and costs

Impact assessments do not distinguish between the relative worth of outcomes, instead treating all programme benefits and all programme costs as equal. Benefits as disparate as increased household consumption and increased participant self-esteem, for example, are treated as equal, as are similar, but clearly distinguishable, benefits, such as full-time and part-time jobs created. In fact, neither all benefits nor all costs are equal. Society attaches greater value to some programme benefits and greater cost to some programme costs than to others. Arguably, therefore, impact assessments should reflect social values via some kind of weighting scheme. Short of an explicit weighting scheme, but still helpful, would be some discussion of relative weights and how they might affect the analysis. Nonetheless, neither weighting schemes nor any discussion of the relative weight of programme outcomes can be found in reviewed impact assessment.

Miscellaneous other methodological shortcomings

Impact assessments are subject to a number of other miscellaneous methodological shortcomings. These methodological shortcomings pertain for the most part to choices made by IA researchers about which information

to report and how to report it. Choices made about the presentation of information can materially influence assessment findings and how findings are interpreted.

Choices made about presentation of information reflect a number of factors, including the skill of the researcher, the objectives of the assessment, subjective decisions about the relevance of information, or the biases of the researcher or the funder. While we cannot assume that choices about presentation of information are necessarily influenced by the biases of the researcher or funder, neither can we assume that biases never figure in the choices. This issue takes on greater relevance when we realise that in many cases impact assessments are funded or conducted by people or organisations that openly advocate microenterprise development. This combined with all the other factors that influence choices about presentation of information suggest that it is useful for policymakers to be familiarised with the ways in which such choices can influence assessment results. Schreiner (2002) catalogues several of the more common examples found in assessments of US microenterprise programmes. The same practices, however, can be found to greater or lesser degrees in other microfinance and microenterprise programme assessments, such that it is useful to summarise Schreiner's arguments below.

Selective Presentation of Programme Outcomes. Impact assessments at times selectively present information in ways that bias the analysis and conclusions. One example is the failure to distinguish between stocks (*e.g.*, loans outstanding or current trainees) and flows (*e.g.*, loans disbursed or people trained). Flows aggregate past performance with current performance, thereby producing a distorted picture of current performance. Flows also exceed stocks at any point in time, such that reporting flows rather than stocks gives a more favourable view of programme performance. Whether it is appropriate to report the one or the other depends on the question asked. Regardless, researchers should report the unit being used and why.

A similar practice is to report aggregate figures as opposed to aggregate figures adjusted for the number of programme participants. A finding, for example, that sales of programme participants totalled $3.5 million sounds more impressive than a finding that average sales per programme participant totalled $12 000.

Another case of selective presentation is the use of half-statistics. Interpretation of findings can be influenced by how they are "spun". For example, reporting that "over one-half" of programme participants increased enterprise returns is a positive spin to the more negative finding that "nearly one-half" of programme participants did not increase enterprise returns. Reporting broad summary statistics when disaggregated statistics are

available is yet one more practice found in impact assessments. For example, the loan repayment rate is a commonly reported measure of portfolio quality; however, this is a broad summary statistic that hides crucial information discernible in far better measures, such as aged portfolio at risk.[17]

A final example of selective presentation is the failure to report programme dropouts. Dropouts are perhaps the simplest, most effective way to gauge whether the programme creates value for participants. It is reasonable to conclude that programmes with high rates of client turnover are less effective at creating value and have less impact than programmes with relatively low rates of client turnover. Moreover, exclusion of dropouts from the impact analysis almost certainly produces systematic overstatement of programme benefits.

Misestimation of Programme Benefits. It is not unusual for impact assessments to misestimate programme benefits. Schreiner points out, for example, that some microenterprise assessments report new businesses starts as if all are attributable directly to programme participation, overlooking that some participants enter the programme with pre-existing enterprises, others start businesses after dropping out, and the majority of new enterprises fail within a few years.

Another example is the misreporting of enterprise income, such as the practice of 1) reporting income levels instead of changes in income, 2) reporting enterprise income instead of enterprise returns, and 3) failing to define income clearly. In the first case, reporting income levels overstates programme benefits unless income was zero prior to joining the programme. In the second case, enterprise income gives a distorted picture of impact because it does not account for returns to time worked and capital invested. Absolute income levels that appear high may not appear so high once adjusted for time worked and capital invested. Finally, lack of clarity about the definition of income allows researchers to report a variety of outcomes as enterprise income, some more favourable than others. A prime example is the practice of reporting unadjusted enterprise revenues as income or as a proxy for income, which inevitably skews findings upwards.

Judging IA methodology

Knowledge of the methodological pitfalls and shortcomings of impact assessments is useful for judging the quality of programme assessments. Before casting judgment, however, it bears repeating that while methodological purity is a worthy ideal, it is rarely achieved in practice. Virtually all impact assessments suffer from one methodological shortcoming or another. Some are blatant, and some require a more careful reading to catch. Some are the result of environmental or resource constraints, some are

the result of subjective choices made by researchers, and some are the result of both.

In practice, most researchers are forced to make concessions of one kind or another to resource or environmental constraints, and they must make subjective choices in how to deal with them. More broadly, all researchers must make a myriad of subjective choices about how to design and implement assessment studies, analyze assessment data, and report assessment findings. While there is little disagreement among researchers on what major methodological issues and pitfalls are or what constitutes ideal methodology, there is substantial disagreement on what constitutes acceptable methodology, what acceptable tradeoffs between rigor and cost are, what methodological concessions one might legitimately make to resource or environmental constraints, or which subjective choices are valid and which not.

Publication in peer-reviewed academic journals by no means implies methodological seal of approval. Editorial boards of academic journals have different methodological standards, as do the reviewers they use. Few published peer-reviewed academic impact assessments satisfy everyone's standards of methodological rigor.[18] But if the methodological gulf between academic researchers is wide, the gulf between academics and practitioners is a positive chasm. Not that practitioners do not care about scientific validity, they do to a degree. It is just a luxury that most of them feel they cannot afford. They are content to let academics hash out methodological niceties, while they concentrate on running programmes. Besides, what manager in any organisation has ever made key programme decisions with a +–.05 degree of certainty?

Given the disagreement about appropriate methodological rigor, policymakers should not be expected to sort out what professional researchers and practitioners cannot. What policymakers can do, however, is to insist on full disclosure of methodological approaches and shortcomings, constraints faced, and tradeoffs and subjective choices made. Full disclosure promotes transparency, and transparency promotes informed judgments about methodological appropriateness and informed interpretation of assessment findings. Full disclosure is an easy, yet critical, objective to which all impact assessments should aspire. As Schreiner (2002) has noted, "The heart of the social-scientific method is not experiments but explicitness." (p. 69)

Recommendations

Taking all of the above into account, the following recommendations are proposed to help policymakers judge the quality/rigor of impact assessments and interpret their findings and to guide their efforts at improving IA practice and usefulness.

Encourage the use of mixed method approaches. The heavy reliance on impact surveys probably reflects as much or more the training, inclinations, and biases of IA researchers than any inherent methodological superiority. Qualitative methods (whether based in the humanities tradition or PLA) have a great deal to offer, particularly in terms producing a deeper and more nuanced understanding of impact. Qualitative methods also can do much to overcome the problems of loan fungibility, and they have been used too infrequently for this purpose. One of the most insightful impact assessments reviewed here was Todd's (1996) ethnographic study of Grameen Bank members; it provided a richness of understanding missing from most survey-based studies. Where time or resources do not permit ethnographic studies of this sort, PLA or other rapid assessment methods can similarly yield insightful and useful information about the process through which impact occurs and how it is manifested at different levels of analysis. The impact survey should remain a primary tool, but it need not be as dominant as it has been to date.

Use valid control groups. If the purpose of a survey-based assessment is to attribute impact to programme participation, a valid control group must be used. It makes little sense to invest scarce public funds in an impact assessment that cannot hope to answer the questions asked of it. At a minimum, a valid control group should consist of microentrepreneurs eligible to join the programme. Pipeline clients satisfy this minimum requirement where binding constraints exist, but efforts should be made to select pipeline clients who have not had the opportunity to join the programme earlier, such as pipeline clients from communities into which the programme has only recently expanded.

Control for selection bias. All reasonable effort should be made to control for all forms of selection bias. Random assignment and random programme placement should be used where possible. Random assignment need not be costly or overly intrusive. Coleman (2001a), for example, estimates that replication of his methodology could easily be implemented at a cost of only $25 000-$50 000, which is close to the minimum range that a high-end impact assessment would cost anyway. The study by Benus *et al.* (1994) demonstrates that random assignment is possible in an OECD context as well. Even were controlling for selection bias to cost more, it is worth an extra increment of spending. Again, it makes little sense to invest money in a programme assessment that can be predicted ahead of time to yield questionable or invalid findings, particularly when valid findings could be produced for only a slightly larger investment. Researchers should also show evidence of controlling for selection bias and be candid in discussing its implications.

Include programme dropouts in the treatment group. The nearly universal exclusion of programme dropouts from treatment groups is a serious methodological shortcoming that has almost assuredly produced systematic overstatement of impact. Including programme dropouts requires additional

investment of money and effort, but it is certainly doable in most cases, and the return in terms of statistical precision is likely to be more than worth the additional investment.

Locate the control group a sufficient distance from the treatment group. To avoid contamination bias, the control group should be located far enough away, within reason, from the control group to minimise the probability of contamination bias.

Allocate sufficient money to conduct methodologically sound impact assessments. If policymakers or funding agencies require programmes to conduct impact assessments, they should allocate sufficient funding to implement them and to implement them in a methodologically sound way. Too often funding agencies expect convincing evidence of impact but do not allocate enough money (or at times any money) to implement a valid impact assessment. All this tends to produce is half-hearted effort and cynicism about impact assessment.

Perform longitudinal assessments where feasible. Cross-sectional assessments are acceptable, but they should be the clear second choice and reserved for those situations in which longitudinal assessments are not feasible.

Agree ahead of time on a clear and complete scope of work with programme management and IA researchers. Methodological problems can arise because of differing expectations or misunderstandings among funding agencies, programme management, and researchers. Thus many methodological problems can probably be avoided if funding agencies work more closely with programme management and researchers to define a scope of work that covers, at a minimum, *a)* the objectives of the assessment, *b)* methodological options consistent with the assessment objectives, *c)* levels of impact to be assessed and corresponding indicators to be used, *d)* field constraints (objectives and methodology may need to be negotiated in light of field constraints), and *e)* reporting requirements (what is to be reported and how).

Perform more rigorous benefit-cost analyses. Whether microenterprise development warrants large expenditures of public money has yet to be determined owing to the dearth of good benefit-cost analyses. While it is helpful to know whether and to what extent microfinance and microenterprise programmes benefit participants and non-participants, this information does not necessarily allow informed choices about whether and to what degree to fund the programmes relative to other policy options. This requires rigorous programme assessments that assess a variety of benefits and costs. It also requires some method to standardise and weight findings.

Schreiner (2002) points out that emphasis on rigorous benefit-cost analysis creates a potential prisoner's dilemma for microfinance and microenterprise programmes in that greater rigor (and disclosure) places it at a disadvantage among policymakers relative to competing policies that do

not adhere to correspondingly high levels of assessment rigor. Thus a subsidiary recommendation is that policy makers hold all programme assessments to high standards of methodological rigor. Regardless, poor assessment methodology in competing social programmes does not relieve microfinance or microenterprise programmes of the ethical responsibility to make reasonable and valid efforts to justify the scarce public funds they receive.

Insist on full and candid disclosure. IA researchers should be held accountable to disclose completely and candidly all information relevant for non-specialists to understand and interpret assessment findings. Things to be disclosed include, at a minimum, *a)* methodologies chosen, why they were chosen, and their implications, *b)* weaknesses of methodologies chosen and their implications, *c)* constraints faced and tradeoffs made and their implications, *d)* subjective decisions made and their implications, *e)* methodological shortcomings and their implications, and *f)* biases or conflicts of interest (real or potential) that may or may not influence methodology, analysis, or interpretation and presentation of findings

Develop IA capacities of practitioner organisations. Any set of recommendations also needs to take into explicit account the drawbacks of the proving approach to impact assessment, principal among them its cost, length, difficulty, and technical requirements, all of which have limited its usefulness to programme management. Given the resource and technical constraints of practitioner organisations, an approach to IA that places scientific validity as the primary criterion is unlikely to be adopted by practitioner organisations to any significant extent. On the other hand, an approach to IA that focuses on the usefulness of IA to programme management, and which can offer reasonable guarantees to management that its benefits in terms of improved programme effectiveness outweigh its costs, is more likely to be adopted by practitioner organisations on a wider scale. This fact needs to be acknowledged at the outset.

Thus if the primary objective of IA is to prove impact, and policymakers are content with assessing a relatively small number of programmes on a sporadic basis, then a reliance on scientific IA is probably appropriate. But if an objective of IA is to improve impact, and policymakers want more or less regular information on programme performance across a wider range of programmes, then a reliance on scientific IA is probably not appropriate. In the latter case, public investment in IA should concentrate increasingly on developing the IA and market research capacity of practitioner organisations, and assisting in the development and implementation of methodologically sound, practical, low-cost, practitioner-friendly, and useful methodologies. This requires in turn that policymakers work with practitioner organisations to understand what their needs and constraints are and work jointly with them to develop a funding and technical assistance policy that passes on critical knowledge and skills so that

programmes can conduct their own assessments in a way that is useful to them and to clients. Moreover, for policymakers wishing to promote programme sustainability, improving impact through IA and market research offers one of the most potentially effective policy approaches.

Notes

1. A high rate of informal sector activity by the poor is a particular characteristic of LDCs.
2. For other reviews of IA methodology, see Gaile and Foster (1996), Khandker (1998), Hulme (2000), Coleman (2001a), and Schreiner (2002).
3. Perhaps the best source for understanding the conceptual foundations of impact assessment and its many complex relationships is the series of conceptual papers commissioned and published by the Assessing the Impact of Microenterprise Services Project (AIMS), funded by the Office for Microenterprise Development at USAID. AIMS publications cover topics such as assessing impact at the enterprise and household levels (Inserra 1996) and at the individual level (Chen 1997), income and assets as impact indicators (Barnes 1996, Little 1997), measuring profits and net worth of microenterprises (Daniels 1999), assessing impacts within a "household economic portfolio" (Chen and Dunn 1996), and microfinance and risk management (Cohen and Sebstad 1999). All AIMS publications can be downloaded at *www.usaidmicro.org*.
4. The terms "scientific method", "humanities tradition", and "PLA" used here to describe IA methodological approaches, as well as the ensuing summaries of each approach, are based on Hulme's (2000) excellent discussion of IA methodologies.
5. Many "participatory" studies are participatory in name only. The term *participation* is fast reaching cliché status – oft used and oft devoid of substantive meaning.
6. SEEP refers to the Small Enterprise Education and Promotion Network. SEEP is a professional network of North American Private Voluntary Organisations engaged in the promotion of microfinance. SEEP worked in conjunction with the AIMS research team to develop the SEEP/AIMS tools.
7. See Nelson *et al*. (2001) for an in-depth description of and implementation instructions for each of the SEEP/AIMS tools.
8. MicroSave Africa, for example, has developed a set of qualitative market research tools using largely PLA methodologies. To date, MicroSave has conducted several training workshops around the globe involving dozens of MFIs. Information on the MicroSave market research tools can be viewed at *www.microsave-africa.com*.
9. For example, a late joiner may only recently have moved into the community, or she may have only recently started a business, or she may not have heard about the programme until recently.
10. An example of how not to select the control group is Kosanovich and Fleck (2002) who selected control group members from among individuals offered enrolment in the microenterprise programme but who declined. This approach virtually guarantees significant selection bias.
11. Pitt and Khandker's approach was criticised by Morduch (1999), who questioned whether the three programmes assessed actually enforced the asset-based eligibility rules, and who found other problems with Pitt and Khandker's

econometric methodology. Pitt's (1999) response to Morduch defended the original approach and purported to demonstrate how it was superior to the alternative methodology proposed by Morduch.

12. Hulme (2000) points out that diversion of loans for consumption may in fact produce higher returns than investment in enterprise assets, for example, if spending on consumption takes the form of investment in human capital (*e.g.*, school fees, health care), replaces borrowing from other sources at a higher cost, or is used to acquire basic needs (*e.g.*, food, medicine) necessary to sustain adequate levels of labour productivity.

13. It would also be useful to include three-year clients and on up in the treatment group; however, in many MFIs, three-and-four year clients on up are hard to come by, owing to high client dropout rates.

14. This author, for example, participated in a longitudinal impact assessment of a microfinance programme in Tanzania that suffered from over 80 per cent panel attrition after just over one year into a two-year study.

15. Woller and Parsons (2002), for example, find that microfinance programmes can contribute from hundreds of thousands to millions of dollars to local economies via direct expenditures and income multipliers. Other impact assessments examining community-level impacts include Coleman (1999, 2001b), Khandker (1996, 2001), Khandker *et al.* (1998), Kilby and D'Zmura (1985), Morduch, (1998), Pitt *et al.* (1999), Schuler and Hashemi (1994), Schuler, Hashemi, and Riley (1997), and Zeller *et al.* (1996).

16. The opportunity cost of labour occurs when new employees leave a previous job and are not replaced, or are replaced by less productive workers.

17. For a good critique that demonstrates how loan repayment rates can be misleading, see Rosenberg (1999).

Appendix

Recommendations for mid-range impact assessments

In April 1997 and April 1998 the Consultative Group to Assist the Poorest (CGAP) conducted two virtual meetings of microfinance experts to develop methodological guidelines for midrange assessments (Cohen and Gail, 1998). They reached consensus on the following guidelines:

1. Use some form of time perspective. Allow enough time for impacts to occur (both in terms of client participation and programme maturity).
2. Use some form of comparison group. Non-clients are preferable to pipeline clients where possible. The higher cost and other limitations of this approach, however, are well-recognised.
3. Tailor assessments to the specific context being studied.
4. Begin with a small set of indicators that have demonstrated validity in previous IA studies and that are relatively easy to collect. Incorporate new indicators as appropriate.
5. Collect baseline indicators, if possible when clients enter the programme. If not possible, use retrospective information.
6. Use interval-level data where possible.
7. Make greater use of IA as a management tool for generating information that is useful for programme improvement; for example, incorporate client satisfaction into IA studies.
8. Employ methods to establish plausible association between programme participation and observed outcomes.
9. Incorporate client satisfaction as part of IA.
10. Use a carefully designed IA that ensures transparency and external review/ oversight.

11. Incorporate plans for IA into programme design and implementation as early as possible.
12. Measure direction of change where exact change cannot be estimated.
13. Build local capacity to conduct IA, both internal and external to the organisation.

References

ASHE, Jeffrey and Lisa PARROTT (2001), *PACT's Women's Empowerment Program in Nepal. A Savings and Literacy Led Alternative to Financial Institution Building*. Washington, DC: PACT.

ASHE, Jeffrey and Michael MACINTYRE (2000), *Working Capital. Building the Grass Roots Economy of Low Income Communities. Part 1: Membership Characteristics and Program Impact*. Cambridge, MA: Working Capital.

BARNES, Carolyn (1996), "Assets and the Impact of Microenterprise Finance Programs". AIMS Paper. Washington, DC: Management Systems International.

BARNES, Carolyn (2001), "Microfinance Program Clients and Impact: An Assessment of Zambuko Trust, Zimbabwe". AIMS Paper. Washington, DC: Management Systems International.

BARNES, Carolyn, Gaile MORRIS and Gary GAILE (1999), "An Assessment of Clients of Microfinance Programs in Uganda". *International Journal of Economic Development*, Vol. 1, No. 1, 80-121.

BENDICK, M. and M.L. EGAN (1987), "Transfer Payment Diversion for Small Business Development: British and French Experience". *Industrial and Labour Relations Review*, Vol. 40, No. 4, 528-542.

BENUS, Jacob M., Michelle WOOD and Neelima GROVER (1994), *A Comparative Analysis of the Washington and Massachusetts UI Self-Employment Demonstrations*. Bethesda, MD: Abt Associates.

BLAIR, Amy Kay and Joyce KLEIN (2001), *Microenterprise as a Welfare to Work Strategy: Client Characteristics*. Washington, DC: The Aspen Institute.

BOLNICK, Bruce R. and Eric R. NELSON (1990), "Evaluating the Economic Impact of a Special Credit Programme: KIK/KMKP in Indonesia". *The Journal of Development Studies*, Vol. 26, No. 2, 299-312.

BUCKLEY, Graeme (1996a), "Financing the Jua Kali Sector in Kenya. The K-Rep Juhudi Scheme and Kenya Industrial Estates Informal Sector Program". In *Finance Against Policy, Volume II: Country Case Studies,* edited by David Hulme and Paul Mosley, 271-332, London: Routledge.

BUCKLEY, Graeme (1996b), "Rural and Agricultural Credit in Malawi. A Study of the Malawi Mudzi Fund and the Smallholder Agricultural Administration". In *Finance Against Policy, Volume II: Country Case Studies*, edited by David Hulme and Paul Mosley, 333-408, London: Routledge.

BUVINIC, Mayra, Marguerite BERGER and Cecelia JARAMILLO (1989), "Impact of a Credit Project for Women and Men Microentrepreneurs in Quito, Ecuador". In *Women's Ventures*, edited by M. Berger and M. Buvinic, 222-246, West Hartford, CT: Kumarian Press.

CHAMBERS, Robert (1997), *Whose Reality Counts? Putting the First Last*. London: IT Publications.

CHEN, Martha Alter (1997), "A Guide for Assessing the Impact of Microenterprise Services at the Individual Level". AIMS Paper. Washington, DC: Management Systems International.

CHEN, Martha Alter and Elizabeth DUNN (1996), "Household Economic Portfolios". AIMS Paper. Washington, DC: Management Systems International.

CHEN, Martha Alter and Donald SNODGRASS (2001), "Managing Resources, Activities, and Risk in Urban India: The Impact of SEWA Bank". AIMS Paper. Washington, DC: Management Systems International.

CHURCHILL, Craig F. (1995), "The Impact of Credit on Informal Sector Enterprises in South Africa: A Study of Get Ahead Foundation's Stokvel Lending Programme". MA Thesis, Worchester, MA: Clark University.

CLARK, Peggy and Tracy HUSTON (1993), *Assisting the Smallest Businesses: Assessing Microenterprise Development as a Strategy for Boosting Poor Communities*. Washington, DC: The Aspen Institute.

CLARK, Peggy, Amy KAYS, Lily ZANDNIAPOUR, Enrique SOTO and Karen DOYLE (1999), Washington DC: The Aspen Institute.

COHEN, Monique (1999), "Opening Up the Impact Assessment Agenda". AIMS Paper. Washington, DC: Management Systems International.

COHEN, Monique and Gary GAILE (1998), "Highlights and Recommendations of the Second Virtual Meeting of the CGAP Working Group on Impact Assessment Methodologies: April 14-28, 1998. Developing Lower Cost Microenterprise Impact Assessment Methodologies for Microenterprise Programs". AIMS Paper. Washington, DC: Management Systems International.

COHEN, Monique and Jennefer SEBSTAD (1999), "Microfinance and Risk Management: A Client Perspective". AIMS Paper. Washington, DC: Management Systems International.

COHEN, Monique and Jennefer SEBSTAD (2000), "Microfinance, Risk Management, and Poverty". AIMS Paper. Washington, DC: Management Systems International.

COLEMAN, Brett E. (1999), "The Impact of Group Lending in Northeast Thailand". *Journal of Development Economics*, Vol. 60, No. 2, 105-142.

COLEMAN, Brett E. (2001a), "Measuring Impact of Microfinance Programs". *ADB Finance for the Poor*, Vol. 2, No. 4, 5-8

COLEMAN, Brett E. (2001b), "Microfinance in Northeast Thailand: Who Benefits and How Much?" Mimeo. Asian Development Bank.

COPESTAKE, James (2000), "Integrating Impact Monitoring and Assessment of Microfinance". *Journal of Development in Practice,* Vol. 10, No. 5, 705-711.

COPESTAKE, James, Sonia BHALOTRA and Susan JOHNSON (2001), "Assessing the Impact of Microcredit: A Zambian Case Study". *The Journal of Development Studies*, Vol. 37, No. 4, 81-100.

CREEVEY, LucY E. (1994), "Summary Report: A UNIFEM Comparative Study Assessing Impacts of Projects to Support Microenterprises".

CREEVEY, Lucy E., Koumakh NDOUR and Abdourahmane THIAM (1995), "Evaluation of the Impacts of PRIDE/VITA (The Guinea Rural Enterprise Development Project)". GEMINI Technical Report, No. 94. Bethesda, MD: Development Alternatives, Inc.

DANIELS, Lisa (1999), "Alternatives for Measuring Profits and New Worth of Microenterprises". 1999. AIMS Paper. Washington, DC: Management Systems International.

DEARDEN, Kirk and NazmuL KHAN (1994), "Assessing the Impact of Women's Savings and Credit Programs on Fertility: A Case from Bangladesh". Save the Children Women/Child Impact Program Monograph No. 2, Wesport: CT.

DIAGNE, Aliou (1998), "Impact of Access to Credit on Income and Food Security in Malawi". FCND Discussion Paper No. 46. Washington, DC: International Food Policy Research Institute.

DUNN, Elizabeth and Arbuckle J. GORDON Jr. (2001), "The Impact of Microcredit: A Case Study from Peru". AIMS Paper. Washington, DC: Management Systems International.

DRURY, David, Stephen WALSH and Marlene STRONG (1994), *Evaluation of the EDWAA Job Creation Demonstration*. Washington, DC: US Department of Labour, Employment and Training Administration.

DUMAS, Colette (2001), "Evaluating the Outcomes of Microenterprise Training for Low Income Women: A Case Study". *Journal of Developmental Entrepreneurship*, Vol. 6, No. 2, 97-128.

ELSE, John F. and Carmel CLAY-THOMPSON, *Refugee Microenterprise Development: Achievements and Lessons Learned*. Iowa City: Institute for Social and Economic Development.

GAILE, Gary L. and Jennifer FOSTER (1996), "Review of Methodological Approaches to the Study of the Impact of Microenterprise Credit Programs". AIMS Paper. Washington, DC: Management Systems International.

GOETZ, Anne Marie and Rina SEN GUPTA (1996), "Who Takes the Credit? Gender, Power, and Control Over Loan Use in Rural Credit Programs in Bangladesh". *World Development*, Vol. 24, No. 1, 45-63.

GUPTA, Surendra and Mario D. DAVALOS (1993), "Midterm Evaluation of the Microenterprise Development Project of Jamaica". GEMINI Technical Report No. 59. Bethesda, MD: Development Alternatives, Inc.

HASHEMI, Syed M., Sidney Ruth SCHULER and Ann P. RILEY (1996), "Rural Credit Programs and Women's Empowerment in Bangladesh". *World Development*, Vol. 24, No. 4, 635-53.

HIMES, Cristina and Lisa J. SERVON (1998), *Measuring Client Success: An Evaluation of ACCION's Impact on Microenterprises in the United States*. Cambridge, MA: ACCION International.

HULME, David (2000), Impact Assessment Methodologies for Microfinance: Theory, Experience, and Better Practice. *World Development*, Vol. 28, No. 1, 79-98.

HULME, David, Richard MONTGOMERY, with Debapriya BHATTACHARYA (1996), "Mutual Finance and the Poor. A Study of the Federation of Thrift and Credit Cooperatives in Sri Lanka (SANASA)". In *Finance Against Policy, Volume II: Country Case Studies*, edited by David Hulme and Paul Mosley, 177-245, London: Routledge.

INSERRA, Anne (1996), "A Review of Approaches for Measurement of Microenterprise and Household Income. AIMS Paper". Washington, DC: Management Systems International.

INSTITUTE FOR SOCIAL AND ECONOMIC DEVELOPMENT (1994), "Rivercities of Iowa/ Illinois Self-Employment (RISE) Final Report". Iowa City: Institute for Social and Economic Development.

KARLAN, Dean and Gwendolyn ALEXANDER (2002), "Credit to the Poor: Examining the Attrition Bias of Cross-Sectional Impact Assessments". Mimeo. Princeton University.

KEVANE, Michael and Bruce WYDICK (2001), "Microenterprise Lending to Female Entrepreneurs: Sacrificing Economic Growth for Poverty Alleviation". *World Development*, Vol. 29, No. 7, 1225-1236.

KHANDKER, Shahidur R. (1996), "Grameen Bank: Impact, Costs, and Program Sustainability". *Asian Development Review*, Vol. 14, No. 1, 97-130.

KHANDKER, Shahidur R. (1998), "Micro-Credit Programme Evaluation: A Critical Review". *IDS Bulletin*, Vol. 29, No. 4, 11-20.

KHANDKER, Shahid (2001), "Does Micro-Finance Really Benefit the Poor: Evidence from Bangladesh". Paper presented at the Asia and Pacific Forum on Poverty: Reforming Policies and Institutions for Poverty Reduction, Manila, February 5-9.

KHANDKER, Shahidur R., Hussain A. SAMAD and Zahed H. KHAN (1998), "Income and Employment Effects of Micro-Credit Programmes: Village-Level Evidence from Bangladesh". *Journal of Development Studies*, Vol. 35, No. 2, 96-124.

KILBY, Peter and David D'ZMURA (1985), "Searching for Benefits". USAID Special Study No. 28. Washington, DC: United States Agency for International Development.

KOSANOVICH, William T. and Heather FLECK (2002), *Comprehensive Assessment of Self-Employment Assistance Programs*. Washington, DC: US Department of Labour.

LAPAR, MA. Lucila A., Douglas H. GRAHAM and Richard L. MEYER (1995), "The Effect of Credit on Output: Are the Sectoral Differences?" Rural Finance Program. Department of Agricultural Economics. Columbus, OH: The Ohio State University.

LAPAR, MA. Lucila A, Douglas H. GRAHAM, Richard L. MEYER and David S. KRAYBILL (1995), "Selectivity Bias in Estimating the Effect of Credit on Output: The Case of Rural Non-Farm Enterprises in the Philippines". Rural Finance Program. Department of Agricultural Economics. Columbus, OH: The Ohio State University.

LITTLE, Peter D. (1997), "Income and Assets as Impact Indicators". AIMS Paper. Washington, DC: Management Systems International.

MCKERNAN, Signe-Mary (1996), "The Impact of Microcredit Programmes on Self-Employment Profits: Do Non-credit Programme Aspects Matter?" *The Review of Economics and Statistics*, Vol. 84, No. 1, 93-115.

MKNELLY, Barbara, Chatree WATETIP, Cheryl A. LASSEN and Christopher DUNFORD (1996), "Preliminary Evidence that Integrated and Financial and Educational Services Can Be Effective Against Hunger and Malnutrition". Research paper No. 2. Davis, CA: Freedom from Hunger.

MKNELLY, Barbara and Christopher DUNFORD (1999a), "Impact of Credit with Education on Mothers and their Young Children's Nutrition: CRECER Credit with Education Program in Bolivia". Research Paper No. 5. Davis, CA: Freedom from Hunger.

MKNELLY, Barbara and Christopher DUNFORD (1999b), "Impact of Credit with Education on Mothers and their Young Children's Nutrition: Lower Pra Rural Bank Credit with Education Program in Ghana". Research Paper No. 4. Davis, CA: Freedom from Hunger.

MONTGOMERY, Richard, Debapriya BHATTACHARYA and David HULME (1996), "Credit for the Poor in Bangladesh. The BRAC Rural Development Programme and the Government Thana Resource Development and Employment Programme". In *Finance Against Poverty, Volume II: Country Case Studies,* edited by David Hulme and Paul Mosley, 94-176, London: Routledge.

MORDUCH, Jonathan (1998), "Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh". Cambridge, MA: Harvard University.

MOSLEY, Paul (1996a), "India. The Regional Rural Banks". In *Finance Against Policy*, Volume II: Country Case Studies, edited by David Hulme and Paul Mosley, 246-270, London: Routledge.

MOSLEY, Paul (1996b), "Indonesia: BKK, KUCK, and the BRI Unit Desa Institutions". *In Finance Against Poverty, Volume II: Country Case Studies*, edited by David Hulme and Paul Mosley, 32-93, London: Routledge.

MOSLEY, Paul (1996c), "Metamorphosis from NGO to Commercial Bank: The Case of Bancosol in Bolivia". In *Finance Against Policy, Volume II: Country Case Studies*, edited by David Hulme and Paul Mosley, 1-31, London: Routledge.

MOSLEY, Paul (2001), "Microfinance and Poverty in Bolivia". *The Journal of Development Studies*, Vol. 37, No. 4, 101-132.

MOSLEY, Paul and David HULME (1998), "Microenterprise Finance: Is There a Conflict Between Growth and Poverty Alleviation?" *World Development*, Vol. 26, No. 5, 783-790.

MT. AUBURN ASSOCIATES (1998), *Evaluation of Working Capital Delaware*. Somerville, MA: Mt. Auburn Associates.

MUSTAFA, SHAMS, ISHRAT ARA, DILRUBA BANU, ALTAF HOSSAIN, AJMAL KUBIR, MOHAMMAD MOHSIN and ABU YUSUF (1995) "Impact Assessment Study of BRAC's Rural Development Programme". Final Report.

NEILL, CAtherine, Mario DAVALOS, Washington KIIRU, M. MANUNDU and JEnnefer SEBSTAD (1994), "The Kenya Rural Enterprise Programme under Cooperative Agreement No. AID-615-0282-A-00-7026-00: A Final Evaluation". GEMINI Technical Report No. 77. Bethesda, MD: Development Alternatives, Inc.

NELSON, Eric (1984), "The Economic Impact of KIK/KMKP Credit on Indonesian Small Enterprises". Draft Memo to Kapala Urusan Kredit Koperasi dan Kredit Kecil. Mimeo.

NELSON, Eric and Bruce R. BOLNICK (1986), "Survey Methods for Assessing Small Credit Programs Evaluating the Economic Impact of KIK/KMPK Credits". Jakarta.

NELSON, Candace, Barbara MKNELLY, Elaine EDGCOMB, Gary GAILE, Carter GARBER, Nancy HORN, Karren LIPPOLD and Brian BEARD (2001), "Learning from Clients: Assessment Tool for Microfinance Practitioners". AIMS Paper. Washington, DC: Management Systems International.

OLDHAM, Linda, Hager EL HADIDID, Hussein TAMA, Michele SULIMAN NASHED, SHERIF EL DEWANY, Mahmoud HUSSEIN, Mohammed ABDEL AZIZ and Mohamed SAKR (1994), "Measuring Socioeconomic Impact of Credit on SMI: Assessment of the Monitoring System used by the Alexandria Businessmen's Association", Egypt. GEMINI Technical Report No. 76. Bethesda, MD: Development Alternatives, Inc.

PARK, Albert and Changqing REN (2001), "Microfinance with Chinese Characteristics". *World Development*, Vol. 29, No. 1, 39-62.

PITT, Mark M. (1999), "Reply to Jonathan Morduch's 'Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh'", Providence, RI: Brown University.

PITT, Mark M. and Shahidur R. KHANDKER (1996), "Household and Intrahousehold Impacts of the Grameen Bank and Similar Targeted Credit Program in Bangladesh". World Bank Discussion Paper No. 320. Washington, DC: World Bank.

PITT, Mark M. and Shahidur R. KHANDKER (1998), "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?" *Journal of Political Economy*, Vol. 106, No. 5, 958-996.

PITT, Mark M., Shahidur KHANDKER, Osman H. CHOWDHURY and Daniel L. MILLIMET (1999), "Credit Programs for the Poor and Health Status of Children in Rural Bangladesh". PSTC Working Paper #98-01. Providence, RI: Brown University.

PITT, Mark M., Shahidur KHANDKER, Signe-Mary MCKERNAN and M. ABDUL LATIF (1999), "Credit Programs for the Poor and Reproductive Behaviour in Low-Income Countries: Are Reported Causal Relationships the Result of Heterogeneity Bias?" *Demography*, Vol. 36, No. 1, 1-21.

PULLEY, Robert V. (1989), *Making the Poor Creditworthy. A Case Study of the Integrated Rural Development Program in India*. Washington, DC: The World Bank.

RAHEIM, Salome (1996), "Microenterprise as an Approach for Promoting Economic Development in Social Work: Lessons from Self-Employment Investment Demonstration". *International Journal of Social Work,* Vol. 39, 69-82.

RAHEIM, Salome and Jason FRIEDMAN (1999), "Microenterprise in the Heartland: Self-Employment as a Self-Sufficiency Strategy for TANF Recipients in Iowa 1993-1998". *Journal of Microfinance*, Vol. 1, No. 1, 66-90.

ROSENBERG, Richard (1999), "Measuring Microcredit Delinquency: Ratios Can Be Harmful to Your Health". Occasional Paper No. 3. Washington, DC: Consultative Group to Assist the Poorest.

SCHREINER, Mark (2002), "Evaluation and Microenterprise Programs in the United States". *Journal of Microfinance*, Vol. 4, No. 2, 67-91.

SCHULER, Sidney Ruth and Syed M. HASHEMI (1994), "Credit Programs, Women's Empowerment, and Contraceptive Use in Rural Bangladesh". *Studies in Family Planning*, Vol. 25, No. 2, 65-76.

SCHULER, Sidney Ruth, Syed M. HASHEMI and Ann P. RILEY (1997), "The Influence of Women's Changing Roles and Status in Bangladesh's Fertility Transition: Evidence from A Study of Credit Programs and Contraceptive Use". *World Development*, Vol. 25, No. 4, 563-575.

SCHULER, Sidney Ruth, Syed M. HASHEMI and Shamsul HUDA BADAL (1998), "Men's Violence Against Women in Rural Bangladesh: Undermined or Exacerbated by Microcredit Programmes?" *Development in Practice*, Vol. 8, No. 2, 148-157.

SEBSTAD, Jennefer (1992), "Get Ahead Foundation Credit Programs in South Africa: The Effects of Loans on Client Enterprises". GEMINI Technical Report, No. 44. Bethesda, MD: Development Alternatives, Inc.

SEBSTAD, Jennefer and Sara LOZA (1993), "Lending and Learning: Formal Banks and Microenterprise in Egypt". Report to the Ford Foundation, Cairo. Washington, DC: Community Economics Corporation.

SEBSTAD, Jennefer and Martin WALSH (1991), "Microenterprise Credit and its Effects in Kenya: An Exploratory Study". Report prepared for USAID AFR/MDI and SandT/WID. Washington, DC: Coopers and Lybrand.

SEKKESAETER, Unni Beate (2002), *Evaluation of Hordaland Network Credit. A Microfinance Programme in Hordaland County, Norway – Part of the "Equal Credit" Project under the EU Receite II Programme*. Bradford, West Yorkshire: Bradford School of Management.

SMITH, Stephen C. (2002), "Village Banking and Maternal and Child Health: Theory and Evidence from Ecuador and Honduras". *World Development*, Vol. 30, No. 4, 707-723.

STEELE, Fiona, SajedA AMIN and Ruchira T. NAVED (2001), "Savings/Credit Group Formation and Change in Contraception". *Demography*, Vol. 38, No. 2, 267-282.

SUTORO, Ann Dunham (1990), "KUPEDES Development Impact Survey: Briefing Booklet". BRI, Planning, Research and Development Department.

THE ROBERTS FOUNDATION (1995), *Self Employment and Very Low-Income Women. A Final Report on the Experience of the San Francisco Homeless Women's Economic Development Project*. San Francisco: The Roberts Foundation.

TODD, Helen (1996), *Women at the Center. Grameen Bank Borrowers after One Decade*. Boulder: Westview Press.

US DEPARTMENT OF HEALTH AND HUMAN SERVICES (1994a), "Capital Opportunities Expansion Program. Human Resource Development Council of District IX, Inc.: Bozeman, Montana" In *Micro Business and Self-Employment: Summary of Final Evaluation Findings from 1990*, 54-83, Washington, DC: US Department of Health and Human Services.

US DEPARTMENT OF HEALTH AND HUMAN SERVICES (1994b), "Micro-Business Development Program. Central Vermont Community Action Council, Inc.: Barre, Vermont". In *Micro Business and Self-Employment: Summary of Final Evaluation Findings from 1990*, 1-19, Washington, DC: US Department of Health and Human Services.

US DEPARTMENT OF HEALTH AND HUMAN SERVICES (1994c), "Microenterprise Development Program (MEDP). Mayor's Office of Community Services: Philadelphia, Pennsylvania". In *Micro Business and Self-Employment: Summary of Final Evaluation Findings from 1990*, 38-53, Washington, DC: US Department of Health and Human Services.

US DEPARTMENT OF HEALTH AND HUMAN SERVICES (1994d), "Small Enterprise and Family Development. Southeast Iowa Community Action: Burlington, Iowa". In *Micro Business and Self-Employment: Summary of Final Evaluation Findings from 1990*, 20-37, Washington, DC: US Department of Health and Human Services.

VENGROFF, Richard and Lucy CREEVEY (1994), *Evaluation of Project Impact: ACEP Component of the Community Enterprise Development Project*. Storrs: The University of Connecticut.

WOLLER, Gary and Robert PARSONS (2002), "Assessing the Community Economic Impact of Microfinance Institutions", *Journal of Developmental Entrepreneurship*, Vol. 7, No. 2, 133-150.

WYDICK, W. BrucE (1999a), "Credit Access, Human Capital, and Class Structure Mobility". *The Journal of Development Studies*, Vol. 55, No. 6, 131-152.

WYDICK, W. Bruce (1999b), "The Effect of Microenterprise Lending on Child Schooling in Guatemala". *Economic Development and Cultural Change*, Vol. 47, No. 4, 853-869.

ZAMAN, Hassan (2001), "Assessing the Poverty and Vulnerability Impact of Micro-Credit in Bangladesh: A Case Study of BRAC". Washington, DC: Consultative Group to Assist the Poorest.

ZELLER, Manfred, Akhter AHMED, Suresh BABU, SumiteR BROCA, Aliou DIAGNE and Manohar SHARMA (1996), "Rural Financial Policies for Food Security of the Poor. Methodologies for a Multicountry Research Project". Washington, DC: International Food Policy Research Institute.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

Chapter 18

An Overview of the Panel Discussion: Evaluating Local Economic and Employment Development

by
Alice Nakamura,
Faculty of Business, University of Alberta,
Edmonton, Alberta, Canada

Alistair Nolan of the OECD/LEED Secretariat introduced the final panel debate. The initial questions he posed were: "What is the current state of government commitment to evaluation? Does the present situation need to be improved on? If so, how?" Nolan said that the debate should focus primarily on practical and policy problems associated with evaluation projects. He said the panelists had been asked to each give opening remarks, and that after that he would open the session for questions. The ensuing debate served to illustrate the importance of a number of the technical and policy issues raised in the other conference presentations.

Edward W. Hill, a professor at Cleveland State University in the United States, led off by outlining what he saw as key themes for the debate:

- The definition of a local economic development policy and its outcomes.
- The differences between process and summative evaluations and ways in which each of these were helpful when running programs *versus* when thinking about the causal structure of programs.

The problems associated with the fact that local evaluations are not usually done at the local level.

Hill pointed out that employment policy and economic development are different. He called attention to the challenge of distinguishing labor policy from labor investment, and development spending activities from financial investment in a community context. He felt that more attention should be paid to whether policies had to do with the demand or the supply side of the economy. He also argued against treating labor policy as simply an instrument of economic development. From an economic development perspective, he saw firm profitability as a problematic measure that primarily reflected accounting and national tax code considerations. He cited the Enron disaster as an indication of some of the shortcomings of focusing on firm profitability. He suggested that firm survival rates and product innovation measures that took account of the product life cycle might provide a better basis for producing indicators that would be helpful for economic development purposes. He argued also that income and employment growth measures should be given more prominence in local economic development planning since these reflect the equilibrium of demand and supply in markets.

The Rt. Hon. Henry B. McLeish, a Member of the Scottish Parliament and former First Minister of Scotland, was the next to speak. He noted what he saw

as an erosion of confidence in evaluation and economic development. He felt too that confusion had developed concerning evaluation outcomes and processes. He urged that greater care be taken to make the process of knowledge transfer simple and transparent. He urged those embarking on pilot projects to keep three objectives in mind. The first was to use the information gained from evaluations to reshape the evaluation projects. The second was to carefully think through how to measure the success of a project at completion. And the third was to find ways to apply tangential project developments to improve other aspects of public policy and processes beyond the intended project objectives.

In closing, McLeish noted that policy makers have an inevitable tendency to try to guard the details of evaluation processes, and that political interface of this sort typically is not as transparent as he feels would be desirable. Also, he said that learning (human capital) is a crucial factor for economic development. He said too that quality of life is a dimension that has not been adequately recognized, and argued that quality of life is not properly reflected by the main measures of economic development that are currently in use.

The next panelist to speak was Stephen Wandner from the US Department of Labor. Wandner stated that he would highlight key distinguishing features of the US employment and training programs. He then went on to say that a distinguishing strength of public programs in the United States is that evaluation is a mandated and integral component. He cited the example of the Language and Workforce Investment Act passed in 1998, which called for the implementation of an evaluation process to properly assess the degree of success achieved in realizing the central objectives of the Act. He said that this evaluation process made use of control groups and a scientific random assignment methodology. He explained that modern evaluations of this sort in the United States were an open process. He explained that this openness included giving the public the right to see everything from the data collected (after suitable measures had been taken to project the privacy of individuals) to the interim evaluation reports to the finalized studies. These materials are available to the public, including, of course, the university research community. He acknowledged that, inevitably, not everyone within the policy and research communities liked all aspects of how the evaluations were conducted or the conclusions drawn in the resulting reports. However, he said that it was accepted in the United States that properly conducted evaluations and open access to the information generated by these is essential for informed debate of the issues facing society and that, in this information age, openness of this sort is both a testament to and an essential aspect of the strength of the US political system.

Alice Nakamura, a professor at the University of Alberta School of Business in Canada, spoke next. She strongly endorsed Wandner's remarks about the

importance of evaluation and open access to evaluation data sources and results. She said that Canada had also had some very favorable experiences with this approach, which she said had been pioneered in Canada in the late 1980s and the early 1990s by Human Resources Development Canada (HRDC). HRDC is the Canadian federal government department responsible for the Canadian social insurance program for unemployed workers (formerly Unemployment Insurance, or UI, and now Employment Insurance, or EI, since the passage of Bill C-12 in 1996). She said that HRDC had led the way, in partnership with Statistics Canada, with the development of data resources for evaluating UI/EI programs and by establishing mechanisms allowing researchers outside the government to have access to these data resources.

Nakamura said that HRDC had also provided intellectual leadership and financial support for the formation of a large network, the Canadian Employment Research Forum (CERF), consisting of university as well as government based researchers with methodological expertise in the areas of program evaluation and employment and earnings analysis. She said that this HRDC initiative had succeeded in redirecting the attention of university based scholars in Canada toward Canadian program evaluation. Before the formation of CERF, Canadian academic researchers interested in program evaluation mostly had worked with US data. She said that, in addition to improving the quality of the information for program development in Canada, the HRDC data and research network initiatives had greatly improved the substantive quality of the information about the Canadian UI/EI program that was being delivered in Canadian university classrooms by professors. Indeed, this HRDC initiative has delivered substantial tangential benefits of the sort that McLeish talked about.

The last of the panelists to speak was Professor Philip Davies of the Cabinet Office in the United Kingdom. He argued that in the United Kingdom there had been something of a renaissance of "evidence-based policy". He said that several recent reports had come out that showed an increase in institutional arrangements and funding for policy evaluations. Davis also mentioned that over the previous five years there had been an enormous proliferation of research in the evaluation industry. He then went on to recommend several improvements that he felt should be made in how evaluation project information was disseminated in non-academic settings. He had suggestions for rewarding employees for following evaluation guidelines on methods for using data that had been accumulated from evaluation projects, and on handling the time pressures that he saw as inevitable for government-funded projects. He also had suggestions for strategically allocating funding for research. He concluded his remarks by noting three orders of integration that he saw as necessary in order for the value of evaluations to policy makers to be maximized. The first of these was

the integration of policy evaluations in the sense of looking at the economic, social, environmental, distributional, and risk assessment outcomes in an evaluation. The second was the integration of methodologies including those for multilevel evaluations incorporating both *supra* and sub-national data. And the third was the integration of policy design and implementation in the evaluation context.

Following the initial remarks of the panelists, the discussion was opened up with questions from both Nolan and the audience. The initial round of questions raised the following issues about the role of a central governing authority:

- Where is the impetus going to come from for improving evaluation at the local level?
- Must it come from the center?
- Must it involve the mandating of evaluations and the right incentives to encourage a higher quality of evaluation?

In response, Wandner gave the example of the Workforce Investment Act. He said that the Department of Labor had met with state officials and had encouraged them to build evaluation capacity. However, he admitted that capacity constraints had continued to be a problem in some states, including a number of the smaller ones. McLeish said he felt that at the local level there was less of an established culture of progressive evaluation. He felt that the culture in many localities needed to be changed so that there was less of a focus on the outcomes of projects and more attention to the processes adopted in these projects and to the innovations involved.

Hill argued that it was important to bear in mind the sources of funding for evaluations. He also suggested that, at the local level, summative evaluations were less important than process evaluations.

Nakamura suggested that there might be a potential business interface for local evaluations. She said that many businesses undertook their own local evaluations on a regular basis, though she acknowledged that businesses seemed to be more interested in process than in summative evaluations. Davies felt that having the same governing party have the ownership of the local development problems and the evaluation evidence helped to align the incentives for evaluators.

In sharing evaluation stories, it became clear that there had been a remarkable range of successful applications in a number of countries.

Finally, attention was paid to how public access to evaluation data and reports could be facilitated while also protecting the privacy of individuals. It seemed clear that there would be ongoing tensions between political instincts and the imperatives of informed program analysis in an information age. It

also seemed clear that, whereas the US seemed to have moved ahead to make public access to evaluation data and reports a priority, there were many other countries where this had not happened.

There was also discussion about finding ways to improve the communication of evaluation information to the mass media. This discussion followed up on Hill's earlier remarks about the importance of approaching different audiences in different ways. Hill added that government departments needed to put in place formal communications strategies from the beginning of evaluation projects.

ISBN 92-64-01708-9
Evaluating Local Economic and Employment Development
How to Assess What Works among Programmes and Policies
© OECD 2004

About the Authors and Contributors

Dr. Timothy Bartik

Timothy J. Bartik is senior economist at the W.E. Upjohn Institute for Employment Research, an independent non-profit and non-partisan research organisation in Kalamazoo, Michigan, USA. He received his Ph.D. in Economics from the University of Wisconsin-Madison in 1982. Dr. Bartik was Assistant Professor of Economics at Vanderbilt University prior to joining the Institute in 1989. At the Institute, Dr. Bartik is responsible for conducting research on state and local economic development policies, local labor markets, and urban poverty problems. He has written two books: "Who Benefits from State and Local Economic Development Policies?" (Upjohn Institute, 1991), and "Jobs for the Poor: Can Labor Demand Policies Help?" (New York: Russell Sage Foundation, 2001). Among his numerous scholarly articles are: "Strategies for Economic Development", in the book "Management Policies in Local Government Finance" (edited by J.R. Aronson and E. Schwartz); "Michigan's Economic Development Policies" (with P. Eisinger and G. Erickcek) in the book "Michigan at the Millennium" (Michigan State University Press, 2003); "Spillover Effects in State Labor Markets of Welfare Reforms", in the Journal of Regional Science in November 2002; "Can Economic Development Programs Be Evaluated? (with R. Bingham) in the book "Dilemmas of Urban Economic Development" (Sage Publications, 1997); and "Who Benefits from Local Job Growth, Migrants or the Original Residents?" in the journal Regional Studies (September 1993).

W. E. Upjohn Institute for Employment Research
300 S. Westnedge Avenue
Kalamazoo, Michigan 49007
USA

Tel.: +1 (616) 385 0433
Fax: +1 (616) 343 3308
E-mail: BARTIK@we.upjohninst.org

* * *

Professor Daniele Bondonio

Daniele Bondonio is Assistant Professor at the *Università del Piemonte Orientale,* Alessandria, Italy. He teaches courses in econometrics and quantitative methods for program evaluation and public policy analysis at the *Università del Piemonte Orientale* and the *Università di Torino*. He received a Ph.D. in public policy analysis and management from *Carnegie Mellon University*, Pittsburgh, USA, where he was visiting scholar in 2001, and a B.A. cum laude in economics from the *Università di Torino*. Professor Bondonio's primary research interests focus on impact evaluations of local economic development and geographically-targeted business incentive programs. His Ph.D. dissertation, which analyzes a number of State Enterprise Zone Programs in the US, received the *National Science Foundation* doctoral dissertation research grant and the *Department of Housing and Urban Development* doctoral dissertation research award. His recent publications focus on impact evaluations of EU business incentive programs in industrially declining areas, investment subsidies targeted to youth-owned firms and methods to comparatively evaluate economic development programs heterogeneously implemented across different regions or states. Professor Bondonio has also produced consulting reports for Italian regional governments and he is part of *The Evaluation Project* in Torino, Italy, aimed at promoting impact evaluation practices in public administration.

Department of Public Policy and Public Choice
Università del Piemonte Orientale
Via Cavour 84
15100 Alessandria
Italy

Tel.: +39 011 533191/0131 283712
Fax: +39 011 5130721
E-mail: daniele.bondonio@sp.unipmn.it

* * *

Dr. Paola Casavola

Paola Casavola received a Law Degree from the University of Naples, Italy, in 1986, a Master of Science in Economics from the London School of Economics in 1989 and a Doctorate in Economics from the University of Naples, Italy, in 1991. From 1991 to 1999 she has worked as an economist in the Research Department (Servizio Studi) of the Central Bank of Italy in Rome and done research mainly on labour market, local development and corporate governance. Since mid-1999 she has worked in the Evaluation Unit (UVAL) of the Department for Development (DPS) of the Italian Ministry of the Economy

(MEF). Within the Evaluation Unit, as coordinator of the Programme Evaluation Area, she is responsible for a set of activities directed at building and disseminating methodological tools for programme evaluation and for the management of evaluation projects. Within the Department for Development she is member of the coordination committee for the Annual Report on Territorial Policies. She also continues her research work on labour market and local development policies.

Dipartimento per le Politiche di Sviluppo e Coesione
Unità di Valutazione
Via Nerva 1
00187 Rome
Italy

Tel.:+39 06 4761 9079
+39 06 4761 9040 (secretary)
Fax: +39 06 4761 9037-47619075
E-mail: paola.casavola@tesoro.it

* * *

Dr. Philip Davies

Philip Davies is currently on secondment from the University of Oxford to the UK Cabinet Office where he is Director of Policy Evaluation in the Strategy Unit. At Oxford Dr. Davies is Director of Social Sciences in the Department for Continuing Education, and is a Fellow of Kellogg College, Oxford. Philip Davies is a graduate of the Universities of London, Oxford, and California, and has held research and teaching appointments in universities in the United Kingdom and the United States. He was also responsible (with colleagues in the University of Oxford Medical School) for developing the University of Oxford Master's Programme in Evidence-Based Health Care. He directed this programme for its first two years of operation. More recently, Philip Davies has been working with colleagues in Britain and America to develop evidence-based public policy. He is on the Steering Committee of the Campbell Collaboration, which prepares, maintains and disseminates systematic reviews of the effects of interventions in education, crime and justice, and social welfare. He is Chair of the Education Group of the Campbell Collaboration. He is also a Visiting Honorary Fellow of the UK Cochrane Centre. Recent book publications include: co-authorship of *Evidence-Based Health Practice: A Primer for Health Professionals* (Edinburgh, Churchill Livingstone, 1999), and a chapter in *What Works? Evidence and Public Policy* (Bristol, Policy Press, 2000). Recent journal publications include: "What is Evidence-Based Education" (*British Journal of Educational Studies*, 47, 2, 108-120); "The Relevance

of Systematic Reviews to Educational Policy and Practice" (*Oxford Review of Education*, 26, 3and4, 365-378); and "The Campbell Collaboration: Does For Public Policy What Cochrane Does For Health" (*British Medical Journal*, 323, 294-295).

Director of Policy Evaluation
The Strategy Unit
UK Cabinet Office
Room 4.37
Admiralty Arch
The Mall
London, SW1A 2WH
United Kingdom

Tel.: +44 207 276 1864
Fax: +44 207 276 1450
E-mail: phil.davies@cabinet-office.x.gsi.gov.uk

* * *

Dr. Randall Eberts

Randall Eberts is Executive Director of the W.E. Upjohn Institute for Employment Research. His research includes the evaluation of employment and training programs. Dr. Eberts' current work includes developing statistical models to help identify the needs of job seekers so that they can be directed more quickly and effectively to services that best meet their needs. He and his colleagues have developed similar models for the State of Michigan's Worker Profiling and Service Referral (WPRS) system and for a pilot project for welfare-to-work participants funded by the US Department of Labor. Mr. Eberts has also prepared reports for the European Commission on the US experience with early identification of worker needs and the potential of service jobs to stimulate economic growth in Europe. He has also partnered with the OECD's LEED Programme to examine the role of local partnerships in workforce development and economic development. He has published extensively in academic journals, including *Review of Economics and Statistics, Journal of Labor Economics, Journal of Human Resources*, and *Economic Inquiry.* He has authored and edited several books. Previous positions include Associate Professor of Economics at the University of Oregon, Visiting Professor at Texas A&M University, Assistant Vice President and Economist at the Federal Reserve Bank of Cleveland, and Senior Staff Economist on the President's Council of Economic Advisers. Randall Eberts received his Ph.D. from Northwestern University.

W.E. Upjohn Institute for Employment Research
300 South Westnedge Avenue
Kalamazoo, MI 49007
USA

Tel.: +1 (616) 343-5541
Fax: +1 (616) 343-3308E-mail:
EBERTS@we.upjohninstitute.org

* * *

Dr. Kris Hallberg

Kris Hallberg is a Lead Economist in the Operations Evaluation Department of the World Bank, an independent unit within the World Bank that reports directly to the Bank's Board of Executive Directors. Since joining the Bank in 1986, Dr. Hallberg has specialized in trade and industrial policy, private sector development, and small enterprise development, working primarily in Latin America as well as Eastern Europe, Central Asia, and East Asia. Currently, she is responsible for evaluating the Bank's private sector development operations. During the past few years, she has worked with the Committee of Donor Agencies for Small Enterprise Development to produce guidelines for donor intervention in business development services for small enterprises, and is a member of the Donor Committee's working group on impact evaluation. During 1991-94, Dr. Hallberg was the head of the World Bank office in Bogota, Colombia. Prior to joining the World Bank she was on the Faculties of Amherst College and Colby College. Dr. Hallberg holds a Ph.D. in Economics from the University of Wisconsin-Madison.

The World Bank
MSN H3-307
1818 H Street, N.W.
Washington, DC 20433
USA

Tel.: +1 (202) 458 5570
Fax: +1 (202) 522 3123
E-mail: khallberg@worldbank.org

* * *

Professor Edward W. (Ned) Hill

Edward W. (Ned) Hill is Professor and Distinguished Scholar of Economic Development at the Maxine Goodman Levin College of Urban Affairs of Cleveland State University. Mr. Hill is also a Non-resident Senior Fellow of the

Center on Urban and Metropolitan Policy of The Brookings Institution. He edited *Economic Development Quarterly* from 1994 to 2004. Edward Hill was awarded the title of Professor and Distinguished Scholar in the fall of 2001. He was appointed a member of the Board of Trustees of the Cleveland Zoological Society in 2000, appointed to the Board of Directors of the Westside Industrial Retention Network (WIRE-Net) in Cleveland, and the Ohio MEMS Society. Governor, Mr. Robert Taft, appointed him to the Urban Revitalization Task Force in the fall of 1999. Edward Hill and Harold Wolman were awarded the Robertson Prize from the editors of *Urban Studies* in 1994. He is the author of two books, co-editor of four books, and author of over 60 articles, book chapters, and columns. His 2001 book, *Ohio's Competitive Advantage: Manufacturing Productivity*, and other articles and commentaries can be downloaded from: *http://urban.csuohio.edu/faculty/ned_hill/site/index.htm*.

College of Urban Affairs
Cleveland State University
Cleveland, OH 44115
USA

Tel.: +1 (216) 687 2174
Fax: +1 (216) 687 9277
E-mail: Ned@urban.csuohio.edu

* * *

Dr. Peter Huber

Peter Huber is a researcher at the Austrian Institute of Economic Research with a specialisation in regional labour market analysis. He previously worked as researcher at the Department of Economics of the Institute for Advanced Studies, Vienna. He has contributed to a number of studies for the European Union and Austrian regional governments focusing on regional labour market policy both in Austria and the accession candidate countries of the European Union. His research stays include the Palacky University in Olomouc, Czech Republic and the Hochschule für Ökonomie in the former GDR.

WIFO
Postfach 91
A-1103 Wien
Austria

Tel.: +43 1 798 2601 404
Fax: +43 1 798 9386
E-mail: Peter.Huber@wifo.ac.atWebsite: *www.wifo.ac.at/*.

* * *

The Rt. Hon. Henry McLeish

Henry McLeish was in elected office for 30 years at local, regional and national levels of government. He is a former First Minister of Scotland and previous to that was Minister for Enterprise and Lifelong Learning in the newly created Scottish Parliament. Prior to this he was Minister for Devolution in the Blair government at Westminster. As Minister for Devolution he was responsible for piloting the legislation for the Scottish Parliament through the House of Commons and chaired the Committee that prepared the blueprint for the working of the new parliament when it was formally opened in July, 1999. This experience has been combined with nearly 13 years of local government and work in Universities both in Scotland and in the USA.

Mr. McLeish, a graduate in urban planning, has an extensive background as politician, practitioner and academic in the social and economic effects of area based policies at local, regional and national levels of government. His particular interests are the role of governance in economic development, the effective control and monitoring of investment, and how evidence based public policy can be improved by effective evaluation. Henry McLeish has considerable experience of the differences in approach to economic development in Scotland, the United Kingdom and Europe. He has also been involved in government visits to China, Japan, Hong Kong, Taiwan and various States in the United States of America to discuss economic policy.

Current interests include the role of governance in leveraging positive economic change; territorial developments and the issue of competitive economic advantage; the new regionalism in Europe and the contribution devolved government can make to economic improvement; and the contribution that learning, the knowledge economy and technology can make to economic development.

49 George Street
Cellardyke, KY10 3AS
United Kingdom

Tel.: +44 1 333 313 080
E-mail: h.b.mcleish.internet.com

* * *

Dr. Hugh Mosley

Hugh Mosley received his Ph.D. in Political Science from Duke University (USA). He is a Senior Research Fellow in the Labour Market and Employment research unit at the Social Science Research Centre Berlin (WZB), where he has worked since 1986. He has done comparative and interdisciplinary research on a wide range of labour market policy topics. His recent work has been on

implementation issues, especially public employment service reforms, and on policy evaluation. He has published widely on these and other topics and has frequently worked as a consultant to the European Commission and other international organisations on labour market issues. Recent publications include: Effizienz der Arbeitämter (Berlin, 2003) co-authored with Holger Schütz and Günther Schmid; Labour Markets, Gender and Institutional Change (Edward Elgar, 2002) co-edited with Jacqueline O'Reilly and Klaus Schömann; "How can active policies be made more effective?" (with Jaap de Koning) in Günther Schmid and Bernard Gazier, The Dynamics of Full Employment. Social Integration by Transitional Labour Markets (Edward Elgar 2002); Labour Market Policy and Unemployment: Impact and Process Evaluations in Selected European Countries (Edward Elgar, 2001) co-edited with Jaap de Koning.

Wissenschaftszentrum Berlin (WZB)
Reichpietschufer 50
D-10785 Berlin
Germany

Tel.: +49 30 25491 112/126
Fax. +49 30 25491 100
E-mail: mosley@wz.berlin.de

* * *

Professor Alice Nakamura

Alice Nakamura is a professor in the School of Business at the University of Alberta. Her PhD is in Economics from the Johns Hopkins University and she also holds an Honorary Doctor of Law from the University of Western Ontario.

She is the current Academic Co-Chair of CERF, the Canadian Employment Research Forum, and a past President of the Canadian Economics Association. She has been an advisor in the areas of employment policy, productivity, performance measurement and evaluation to governments in Canada, the United States and elsewhere. She is a member of multiple advisory boards for Statistics Canada. Her main research areas and her publications are in labour economics, immigration, econometrics and productivity measurement.

Alice Nakamura is also president of the institute that founded *www.CareerOwl.ca*, an e-recruiting service started by Canadian university faculty members to help reduce the costs to Canadian business of finding the talent they need, and that they helped to train through their tax support of the universities of Canada.

Alberta School of Business
3-23 Business Building
University of Alberta
Edmonton, Alberta
T6G 2R6
Canada

Tel.: +1 780 492 5824
+1 780 492 2457 – secretary
Fax: +1 780 492 9924 or 492 3325
E-mail: alice.nakamura@ualberta.ca

* * *

Mr. Alistair Nolan

Alistair Nolan has worked with the OECD since July 1997, specialising in all aspects of public policy towards entrepreneurship, with a focus on the links between firm creation and the development of local and regional economies. Mr.Nolan played a key role in the preparation of the OECD's 1998 flagship publication *Fostering Entrepreneurship* and was also responsible for two OECD books on business incubation: *Business Incubation: International Case Studies* (1999) and *Good Practice in Business Incubation* (2000). He has also been responsible for OECD policy recommendations on business networks and enterprise clusters. He is the author of the 2003 OECD book *Entrepreneurship and Local Economic Development*, which reviews knowledge in the field and sets out detailed programme and policy guidance for central and local governments. Until January 2004 Mr. Nolan was also responsible for the OECD LEED Programme's work on evaluation methods and practice. He is currently co-managing the preparatory phases of a possible OECD-wide quantitative assessment of adult skills, aimed at shedding light on a broad array of macro- and micro-economic policy concerns. Prior to joining the OECD he worked as one of a small group of staff responsible for monitoring and evaluating the technical assistance programme of the United Nations Industrial Development Organisation. In this context he was responsible for evaluating projects and programmes in fields ranging from training to technology transfer, environmentally clean production and investment promotion. Over a number of years with UNIDO he occupied posts in research, policy and the design of technical co-operation. Mr. Nolan holds a M.Phil. from Cambridge University in the Economics and Politics of Development, as well as post-graduate qualifications in corporate finance and financial economics, as well as studies in environmental economics and project finance. He is registered on the Ph.D. in Economics at Cambridge University.

OECD
2, rue André-Pascal
75775 Paris Cedex 16
France

Tel.: +33 1 45 24 1386
E-mail: Alistair.Nolan@OECD.org

* * *

Dr. Eric S. Oldsman

Eric Oldsman is the founder and President of Nexus Associates, Inc. He has directed numerous projects for a broad range of clients, including federal and state agencies, multilateral institutions, not-for-profit organisations, and leading private corporations in Latin America, Asia, Europe, and the United States. Many of these assignments have focused on the design, management and evaluation of economic development initiatives. In recent years, Dr. Oldsman has been asked to evaluate programs such as the Ben Franklin Technology Partnership (United States), Enterprise Development Centers (Argentina), Industrial Integration Program (Mexico), Mekong Project Development Facility (Vietnam), NIST Manufacturing Extension Partnership (United States), Robert C. Byrd Institute United States), and the Thailand Productivity Institute (Thailand). Evaluations have focused on questions related to organisational development, throughput, operating efficiency, financial self-sufficiency and effectiveness. Various techniques have been used in gauging impacts, including theory-based case studies, customer surveys, and quasi-experimental designs. Prior to founding Nexus Associates in 1991, Dr. Oldsman spent seven years as a senior consultant with Arthur D. Little, Inc. Before that he was on the staff of PACT, Inc. where he was responsible for project monitoring and evaluation of community-based development programs in Africa and Latin America. He holds a Ph.D. in Public Policy from Harvard University and a B.A. in Economics from Brown University.

Nexus Associates, Inc.
68 Leonard Street
Belmont, MA 02478
USA

Tel.: +1 (617) 489 0311
E-mail: oldsman@nexus-associates.com
Website: *www. nexus-associates.com*.

* * *

Dr. Christopher J. O'Leary

Christopher J. O'Leary is a senior economist at the W.E. Upjohn Institute for Employment Research. His research on unemployment insurance has examined reemployment bonuses, profiling, benefit adequacy, and experience rating. He has evaluated training, wage subsidies, public works, self-employment, and employment service programs for labor ministries in the transition countries of Hungary, Poland, and China. For the US Department of Labor he is working with Randall Eberts to develop a frontline decision support system for one-stop career centers under the Workforce Investment Act. His research has also been sponsored by the World Bank, the International Labor Office, and Human Resources Development Canada. His papers have appeared in *Journal of Human Resources, Journal of Policy Analysis and Management, International Labour Review, New England Economic Review, Economics of Transition,* and *Applied Economics*. Christopher O'Leary completed undergraduate studies at the University of Massachusetts at Amherst and earned a doctorate in economics from the University of Arizona. In 1999 he was elected to the National Academy of Social Insurance.

W.E. Upjohn Institute for Employment Research
300 South Westnedge Avenue
Kalamazoo, MI 49007
USA

Tel.: +1 (616) 343-5541
Fax: +1 (616) 343-3308
E-mail: OLEARY@we.upjohninst.org

* * *

Dr. Jonathan Potter

Jonathan Potter has worked as a senior economist in the OECD Local Economic and Employment Development Programme (LEED) since 1997, and is responsible for LEED activities on entrepreneurship and evaluation. He has authored and edited several OECD books, including *Devolution and Globalisation – Implications for Local Decision-makers; Global Knowledge Flows and Economic Development* and *The Local Dimension of Welfare-to-Work*. He also manages two series of review studies, the OECD Local Entrepreneurship Reviews and a series on Foreign Direct Investment and Local Development. Dr. Potter was previously Senior Consultant at PA Consulting Group in the UK, specialising in public policy evaluation. He has undertaken numerous evaluations for the European Commission, central government and local and regional development agencies and advised on the development of evaluation

methodologies. Dr. Potter holds a Ph.D. in Economics from Cambridge University.

OECD
2, rue André-Pascal
75775 Paris Cedex 16
France

Tel.: +33 1 45 24 89 77
Fax: +33 1 45 24 16 68
E-mail: Jonathan.Potter@OECD.org

* * *

Professor Brian Robson

Brian Robson is Director of the Centre for Urban Policy Studies (CUPS) at Manchester University. He has been Professor of Geography at Manchester since 1977 and before that taught at Cambridge University. As Director of CUPS, he has undertaken numerous research contracts for central government departments, local authorities, and other funders, as well as academic projects for the Economic and Social Research Council. His research has focused on three main areas: evaluating the impact of urban policy initiatives; developing measures of deprivation and social exclusion; and exploring the genesis of regional inequalities. He has published 8 books and major reports and well over 100 academic articles. He has played a leading role in advising government on many of the recent policy initiatives in urban regeneration: the development of the Single Regeneration Budget, Urban Development Corporations, Urban Regeneration Companies, and currently the Housing Market Renewal Fund. He was a member of the fiscal working group of Lord Rogers' Urban Task Force, and a member of the government's Urban Sounding Board. In 2000 he was awarded the Royal Geographical Society's Gold Medal for his contributions to public policy and was recently elected as an Honorary Member of the Royal Town Planning Institute.

Department of Geography
Manchester University
Manchester, M13 9PL
United Kingdom

Tel.: +44 161 275 3639
Fax: +44 161 275 7878
E-mail: Brian.Robson@man.ac.uk

* * *

Professor Jeffrey A. Smith

Jeffrey Smith is Professor of Economics at the University of Maryland. He received his Ph.D. in Economics from the University of Chicago in 1996 and joined the Maryland faculty in 2001. Prior to coming to Maryland he was Associate Professor and CIBC Chair in Human Capital and Productivity at the University of Western Ontario in London, Canada. He received the 1997 Polanyi prize from the Province of Ontario, which is awarded each year to an outstanding young economist in Ontario. His research centers on methods for the evaluation of social programs such as job training for the disadvantaged. He has also written papers examining the labor market effects of university quality and the use of statistical treatment rules to assign persons to government programs. Recent publications include "Substitution and Dropout Bias in Social Experiments: A Study of An Influential Social Experiment" (with James Heckman, Neil Hohmann and Michael Khoo), *Quarterly Journal of Economics* 2000, "The Economics and Econometrics of Active Labor Market Programmes" (with James Heckman and Robert LaLonde) in the *Handbook of Labor Economics, Volume 3A* 1999, and "The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme: Implications for Simple Programme Evaluation Strategies" (with James Heckman), *Economic Journal* 1999 (winner of Royal Economic Society prize for an article in *the Economic Journal* in 1999).

Department of Economics
University of Maryland
3105 Tydings Hall
College Park, MD 20742-7211
USA

Tel.: +1 (301) 405 3532
Fax: +1 (301) 405 3542
E-mail: smith@econ.umd.edu
Website: *www.bsos.umd.edu/econ/faculty/Smith.htm*.

* * *

Dr. George I. Treyz

George I. Treyz has a Ph.D. from Cornell University and a B.A. from Princeton. He is Professor Emeritus in Economics at the University of Massachusetts at Amherst. He is author of the book *Regional Economic Modeling* and is author or co-author of over 25 papers published in professional journals, including the *American Economic Review*, *The Review of Economic Statistics*, and the *International Regional Science Review*. In 1980, Dr. Treyz founded Regional Economic Models, Inc. (REMI) to develop and implement

regional macro-economic models for public and private clients. These models are used for forecasting and policy analysis by major US cities, a large majority of state governments, federal government agencies, universities, and both large and small area government agencies. A research team currently led by Drs. George and Frederick Treyz (Ph.D. and CEO) has recently completed the latest REMI Policy Insight® based in part on a REMI prototype set forth in the November 2000 issue of the *Journal of Regional Science*. It is designed for and in use in areas of varying sizes in the United States and the European Union.

President
Regional Economic Models, Inc.
306 Lincoln Avenue
Amherst, MA 01002
USA

Tel.: +1 (413) 549 1169
Fax: +1 (413) 549-1038
E-mail: georgetreyz@remi.com
Website: *www.remi.com*.

* * *

Professor Robert Walker

Robert Walker is Professor of Social Policy at the University of Nottingham and a Research Fellow at the Institute for Fiscal Studies. He is also a Fellow of the Royal Society of Arts and was formerly Director of the Centre for Research in Social Policy at Loughborough University.

Keen that high quality research should be used to inform the political process and to improve policy with the goal of enhancing all our lives, he undertakes research relevant to the development of welfare policies in Britain and other advanced industrial societies. He also engages in dialogue with policy makers and others wanting to use or support research to bring about positive change.

Robert Walker has conducted research for government departments and international bodies continuously since 1983 and led major evaluations of a number of UK polices, including New Deal for Disabled People, Jobseeker's Allowance and the Social Fund. He has recently been involved in a meta-analysis of US welfare-to-work programmes. His special interests include unemployment and employment progression, poverty and poverty dynamics, social exclusion, children's aspirations, family dynamics and household budgeting strategies. He has published 17 books and over 40 research reports.

Professor of Social Policy
Department of Sociology and Social Policy
University of Nottingham
University Park
Nottingham, NG7 2RD
United Kingdom

Tel.: +44 115 951 4546
Fax: +44 115 951 5232
E-mail: Robert.Walker@nottingham.ac.uk

* * *

Dr. Stephen A. Wandner

Stephen Wandner has been Director of Research and Demonstrations for the US Department of Labor's Employment and Training Administration since January 1997. He has initiated and directed a large number research and demonstration projects dealing with the Workforce Investment Act/training programs, Wagner-Peyser/public labor exchange programs, youth programs, and unemployment insurance. A new project he is initiating will provide microenterprise training, technical assistance, transfer payments and loans to a wide range of labor force participants in three states. He has recently edited and authored a recently published book, *Targeting Employment Services*.

During 1996 he was a visiting senior researcher at the Urban Institute where he completed editing and writing *Unemployment Insurance in the United States: Analysis of Policy Issues (1997)*, and wrote a number of journal articles.

Until 1996, he served as Acting Director and Deputy Director of the Office of Legislation and Actuarial Services of the Unemployment Insurance Service.

He also served as the senior researcher for the Unemployment Insurance Service. This research effort included directing eight large-scale experiments that provided reemployment assistance to dislocated unemployment insurance recipients. The evaluation of one of these projects – a New Jersey demonstration project – provided the basis of Federal legislation enacted in 1993 – the Worker Profiling and Reemployment Services initiative. Another evaluation of two self-employment assistance programs resulted in Federal legislation enacting a Self-Employment Assistance program.

He started his Federal government career with the Unemployment Insurance Service as an actuary and supervisory actuary. He has also worked as a policy analyst for the Department of Commerce. He received his Ph.D. in Economics from Indiana University.

US Department of Labor
200 Constitution Avenue, NW
Washington, DC 20210
USA

Tel.: +1 (202) 693 3663
Fax: +1 (202) 219 9074
E-mail: WANDNER.Stephen@dol.gov

* * *

Ms. Andrea Westall

Andrea Westall is Deputy Director of the New Economics Foundation. She was previously Director of the Policy Unit in the Foundation for Entrepreneurial Management at the London Business School and prior to that a senior research fellow at the Institute for Public Policy Research. In 2001, she wrote a report on social enterprise – *Value-Led Market-Driven: Social Enterprise Solutions to Public Policy Goals* which set out the ways in which social enterprise meet a range of public interest outcomes such as employment, health, financial exclusion and market-creation. She has subsequently been involved in the development of this sector through work in the East Midlands and North West with the Regional Development Agencies, the UK Government's DTI Social Enterprise Unit, and the UK Social Enterprise Coalition. She is currently exploring impact measurement for social enterprises through the EU-funded EQUAL project. Andrea has also been involved with the Performance and Innovation Unit of the UK Cabinet Office review of voluntary sector reform, and ongoing research into income generation and social enterprise by the Charities Aid Foundation and by the National Council for Voluntary Organisations. She has explored the role of enterprise creation and support in regeneration activity, working with the Small Business Service (part of the UK Department for Trade and Industry) as well as private and non-profit sector partners. Her other research and policy interests include public sector reform, corporate responsibility and science policy.

New Economics Foundation
3 Jonathan Street
London, SE11 5NH
United Kingdom

Tel.: +44 207 820 6322
Fax: +44 207 820 6301
E-mail: andrea.westall@neweconomics.org
Website: *www.neweconomics.org*.

* * *

Dr. Robert A. Wilson

Robert Wilson is a Principal Research Fellow in the Institute for Employment Research at the University of Warwick in the UK. The Institute is one of Europe's leading centres for research in the labour market field and it has established an international reputation in the areas of occupational change and skill development. Robert Wilson leads the Institute's labour market forecasting work, although he has researched and published on many other aspects of labour market behaviour. He has a strong interest in technological and structural change and its impact on the labour market. His research has also included both sectoral studies, ranging from engineering through construction to health services, and analyses of specific occupational groups such as professional scientists and engineers. The latter includes the study of the role of such personnel within organisations as well as the education, training and employment situations affecting these occupations overall. As well as editing publications relating to employment forecasts such as *Working Futures* (published by the Sector Skills Development Agency), *Projections of Occupations and Qualifications* (published by the Department of Educations and Skills) and the Institute's *Review of the Economy and Employment*, he has written and edited a number of books including *Employment Forecasting in the Construction Industry; The National Health Service and the Labour Market*; *Technical Change: The Role of Scientists and Engineers;* and *Research and Development Statistics*. Amongst his professional responsibilities, he has been a member of the Medical Workforce Standing Advisory Committee and the Skills Task Force Research Group.

Institute for Employment Research
University of Warwick
Coventry, CV4 7AL
United Kingdom

Tel.: +44 24 7652 4127
Fax: +44 24 7652 4241
E-mail: r.a.wilson@.warwick.ac.uk

* * *

Professor Gary Woller

Gary Woller is Associate Professor of Public Management at the Romney Institute of Public Management, Marriott School, Brigham Young University, where he teaches courses in International Development Management and Public Policy. He has published numerous articles in academic and practitioner journals on international development and microfinance. In addition, he has served as consultant for numerous microfinance institutions,

development banks, international development NGOs, and development consulting firms in the areas of impact assessment and market research. Currently Mr. Woller also serves as chief consultant to the Small Enterprise Education and Promotion Network (SEEP), an international professional association based in Washington, DC. In this capacity, he has organized training in impact assessment and market research tools, coordinated research projects under a grant from the Ford Foundation, and interacted with dozens of microfinance institutions for the purpose of researching and developing institutional capacity in impact assessment and market research. Gary Woller is the co-founder and editor of the Journal of Microfinance. He holds a Ph.D. from the University of Rochester and an MBA and BA from Brigham Young University.

Romney Institute of Public Management
766 TNRB
Brigham Young University
Provo, UT 84602
USA

Tel.: +1 (801) 378-4221
E-mail: wollerg.@yahoo.com

* * *

Dr. Ging Wong

Ging Wong is the Director-General of Management, Regional and Correspondence Services at Canadian Heritage. Prior to this, he was Director of Policy Capacity at the Policy Research Initiative, Privy Council Office, Government of Canada, where he had a primary responsibility for building policy research, functional capacity and networks in support of horizontal policy priorities for the Government of Canada. He was on assignment from Human Resources Development Canada where he was Director of Strategic Evaluation and Monitoring, with a mandate for evaluating social programs that constituted 46 per cent of the federal budget. This included evaluations of public pensions, labour standards and the evaluation and monitoring of Employment Insurance reform impacts on individuals, communities and the economy that was required by legislation to report annually to Parliament for the years 1997-2001. He has made a significant contribution to the practice of evaluation in Canada by integrating evaluation with strategic policy during his tenure through a major, innovative review of unemployment insurance that informed policy discussion and helped to shape legislative reforms. Previously, he initiated sectoral councils for human resources development as Director of Sector Studies and was heavily implicated in formulating Canada's Labour Force Development Strategy reform as Chief of Employment Policy.

Ging Wong is a founder and the past government co-chair of the Canadian Employment Research Forum (CERF). Elected to the present executive of the Canadian Economics Association, he was also an executive member of two major research networks: the Canadian International Labour Network (CILN) and the Western Research Network on Education and Training (WRNET). Ging Wong has been a longstanding Canadian delegate to the Directing Committee of the OECD Local Economic and Employment Development (LEED) Program.

Prior to his recruitment to the Canadian federal government by the External Affairs Department, Ging Wong was an assistant professor at the University of Calgary. He did his honours undergraduate and graduate studies at the University of Calgary, and post-graduate studies at Oxford University, reading labour economics and industrial relations. He maintains his research outreach as an Adjunct Professor, Faculty of Business and Management, University of Alberta and was appointed to a similar position with Wuhan University, China.

Canadian Heritage
25 Eddy Street, 11th Floor
Postal Locator: 25-11-0
Gatineau, Quebec
Canada K1A 0M5

Tel.: +1 (819) 952 3514
Fax: +1 (819) 956 3645
E-mail: ggwong@sympatico.ca

OECD PUBLICATIONS, 2, rue André-Pascal, 75775 PARIS CEDEX 16
PRINTED IN FRANCE
(80 2004 03 1 P) ISBN 92-64-01708-9 – No. 53591 2004