ADVANCES IN ACCOUNTING BEHAVIORAL RESEARCH

VOLUME 5

VICKY ARNOLD
Editor

JAI

LIST OF CONTRIBUTORS

Jesse D. Beeler	Else School of Management Millsaps College, USA
Jason Chong	PricewaterhouseCoopers, Australia
B. Douglas Clinton	Department of Accountancy, College of Business, Northern Illinois University, USA
Dann G. Fisher	Department of Accounting Kansas State University, USA
James E. Hunton	Accountancy Department Bentley College, USA
Carol Jessup	Department of Accountancy University of Illinois at Springfield, USA
Khondkar E. Karim	College of Business Rochester Institute of Technology, USA
Chong M. Lau	Department of Accounting and Finance The University of Western Australia, Australia
Michele Martinez	School of Accountancy University of South Florida, USA
Cindy Moeckel	School of Accountancy and Information Management, Arizona State University, USA
Richard I. Newmark	Department of Accounting University of Northern Colorado, USA
Hossein Nouri	School of Business, The College of New Jersey, USA

R. David Plumlee — School of Accounting & Information Systems, University of Utah, USA

Jacob M. Rose — Department of Accounting and Business Law, Montana State University, USA

John T. Sweeney — School of Accounting, Information Systems & Business Law, Washington State University, USA

Brad Tuttle — Moore School of Business, University of South Carolina, USA

John G. Wermert — College of Business, Middle Tennessee State University, USA

Patrick Wheeler — School of Accountancy, University of Missouri-Columbia, USA

REVIEWER ACKNOWLEDGMENTS

The Editor and Associate Editors at AABR would like to thank the many excellent reviewers who have volunteered their time and expertise to make this an outstanding publication. Publishing quality papers in a timely manner would not be possible without their efforts.

George Aldhizer
University of Northern Kentucky, USA

John C. Anderson
San Diego State University, USA

C. Richard Baker
University of Massachusetts-Dartmouth, USA

Jesse D. Beeler
Millsaps College, USA

Richard A. Bernardi
Naval War College, USA

James Bierstaker
University of Massachusetts-Boston, USA

Dennis M. Bline
Bryant College, USA

Richard G. Brody
University of New Haven, USA

Gregory A. Carnes
Northern Illinois University, USA

Robert H. Chenhall
Monash University, Australia

B. Douglas Clinton
Northern Illinois University, USA

Robert Cochran
University of Richmond, USA

Roger Debreceny
Nanyang Technological University, Singapore

William N. Dilla
Iowa State University, USA

Alan S. Dunk
University of Tasmania, Australia

Christine Earley
University of Connecticut, USA

Diana R. Franz
University of Toledo, USA

Dipankar Ghosh
University of Oklahoma, USA

Jennifer D. Goodwin
University of Queensland, Australia

Severin V. Grabski
Michigan State University, USA

Joanne P. Healy
Kent State University, USA

Heather M. Hermanson
Kennesaw State University, USA

Karen L. Hooks
Florida Atlantic University, USA

James E. Hunton
Bentley College, USA

Donald R. Jones
Texas Tech University, USA

Kathryn Kadous
University of Washington, USA

Carol A. Knapp
University of Central Oklahoma, USA

Teresa Libby
Wilfred Laurier University, Canada

Alan T. Lord
Bowling Green State University, USA

Timothy J. Louwers
Louisiana State University, USA

Peter F. Luckett
University of New South Wales, Australia

R. Murray Lindsay
University of Saskatchewan, Canada

Lokman Mia
Griffith University – Gold Coast, Australia

William R. Pasewark
Texas Tech University, USA

Robert J. Parker
University of South Florida, USA

Peter John Poznanski
Cleveland State University, USA

Robin R. Radtke
University of Texas San Antonio, USA

John Reisch
East Carolina University, USA

Robin W. Roberts
University of Central Florida, USA

Jacob M. Rose
Montana State University, USA

Andrew J. Rosman
University of Connecticut, USA

Axel K-D Schulz
University of Melbourne, Australia

Steve G. Sutton
University of Connecticut, USA

John T. Sweeney
Washington State University, USA

Linda Thorne
York University, Canada

Sandra Vera-Munoz
University of Notre Dame, USA

Patrick Wheeler
University of Missouri-Columbia, USA

Stephen W. Wheeler
University of the Pacific, USA

Stacey M. Whitecotton
Arizona State University, USA

George R. Young II
Florida Atlantic University, USA

EDITOR'S COMMENTS

Welcome to Volume 5 of *Advances in Accounting Behavioral Research.* As the new editor, I am very pleased with both the quantity and quality of submissions over the past 18 months. I believe that Jim Hunton, the previous editor, did a wonderful job of establishing the journal and set a very high standard for me to meet. I hope that I am able to continue the precedent he set and am able to publish very high quality manuscripts that significantly contribute to accounting research. Based on the manuscripts published in this volume, I believe that the quality of behavioral research has continued to rise. This volume contains papers on a variety of topics that touch on nearly all areas of behavioral research. I hope that you find these papers interesting and useful in your research pursuits.

The philosophy that Jim established for the journal continues to be the same and is repeated in the section entitled "Editorial Policy and Submission Guidelines" which is included after these comments. In addition, I am expanding to include a section for literature reviews in future volumes. I believe that literature reviews that synthesize and summarize a stream of research are extremely beneficial especially when those reviews result in frameworks that provide organization to a body of research. Comprehensive literature reviews can help to identify what is known and what is not known about a particular area and guide future research that may be needed. They are also beneficial when research results have been mixed and findings do not necessarily support a single result. I strongly encourage new Ph.D.s to consider whether the literature review section of their dissertation might be appropriate as a basis for writing a manuscript for consideration at *AABR.*

As a final note, I would like to thank Jim for his assistance in assembling this volume. In transitioning the editorship, he continued to handle the papers that were previously submitted to him. As a result, Jim accepted three of the manuscripts contained in this volume; and I greatly appreciate his help.

Vicky Arnold
Editor

EDITORIAL POLICY AND SUBMISSION GUIDELINES

Advances in Accounting Behavioral Research (AABR) publishes articles encompassing all areas of accounting that incorporate theory from and contribute new knowledge and understanding to the fields of applied psychology, sociology, management science, and economics. The journal is primarily devoted to original empirical investigations; however, literature review papers, theoretical analyses, and methodological contributions are welcome. AABR is receptive to replication studies, provided they investigate important issues and are concisely written. The journal especially welcomes manuscripts that integrate accounting issues with organizational behavior, human judgment/decision making, and cognitive psychology.

Manuscripts will be blind-reviewed by two reviewers and an associate editor. The recommendations of the reviewers and associate editor will be used to determine whether to accept the paper as is, accept the paper with minor revisions, reject the paper or to invite the authors to revise and resubmit the paper.

Manuscript Submission

Manuscripts should be forwarded to the editor, Vicky Arnold, at *VArnold@sba.uconn.edu* via e-mail. All text, tables, and figures should be incorporated into a Word document prior to submission. The manuscript should also include a title page containing the name and address of all authors and a concise abstract. Also, include a separate Word document with any experimental materials or survey instruments. If you are unable to submit electronically, please forward the manuscript along with the experimental materials to the following address:

Vicky Arnold, Editor
Advances in Accounting Behavioral Research
Department of Accounting U41A
School of Business
University of Connecticut
Storrs, CT 06269-2041

References should follow the APA (American Psychological Association) standard. References should be indicated by giving (in parentheses) the author's name followed by the date of the journal or book; or with the date in parentheses, as in 'suggested by Earley (2000)'.

In the text, use the form Rosman et al. (1995) where there are more than two authors, but list all authors in the references. Quotations of more than one line of text from cited works should be indented and citation should include the page number of the quotation; e.g. (Dunbar, 2001, p. 56).

Citations for all articles referenced in the text of the manuscript should be shown in alphabetical order in the Reference list at the end of the manuscript. Only articles referenced in the text should be included in the Reference list. Format for references is as follows:

For journals:
Dunn, C. L., & Gerard, G. J. (2001). Auditor efficiency and effectiveness with diagrammatic and linguistic conceptual model representations. *International Journal of Accounting Information Systems, 2*(3), 1-40.

For books:
Ashton, R. H., & Ashton, A. H. (1995). *Judgment and decision-making research in accounting and auditing*. New York, NY: Cambridge University Press.

For a thesis:
Smedley, G. A. (2001). *The effects of optimization on cognitive skill acquisition from intelligent decision aids.* Unpublished doctoral dissertation, University.

For a working paper:
Thorne, L., Massey, D. W., & Magnan, M. (2000). Insights into selection-socialization in the audit profession: An examination of the moral reasoning of public accountants in the United States and Canada. Working paper: York University, North York, Ontario.

For papers from conference proceedings, chapters from book etc.:
Messier, W. F. (1995). Research in and development of audit decision aids. In: R. H. Ashton & A. H. Ashton (Ed.), *Judgment and Decision Making in Accounting and Auditing* (pp. 207–230). New York: Cambridge University Press.

THE IMPACT OF DIGITAL TECHNOLOGY ON ACCOUNTING BEHAVIORAL RESEARCH

James E. Hunton

ABSTRACT

This article reflects on recent dramatic changes that have forever reshaped accounting practice and research. Specifically, the digital revolution, which is still unfolding before our eyes, has changed the very nature of work for accountants and forced researchers and practitioners alike to struggle with a host of new threats and opportunities facing the profession. Accounting behavioral researchers in particular must deal with a wide array of human-to-computer, information processing and decision-making issues that did not exist a mere two decades ago. We academics can put our collective heads in the sand, as some have already chosen to do, or we can view these turbulent times in the global business community as exciting, provoking and enlightening. Herein, I attempt to address many ways in which psychology-based researchers can engage in meaningful studies aimed at raising the accounting profession to a new level of quality and relevance by incorporating the impact of information and communication technologies on human attitude, cognition and performance.

Advances in Accounting Behavioral Research, Volume 5, pages 3–17.

ISBN: 0-7623-0953-9

INTRODUCTION

Many of us look back with fondness on the days when we were children and school was out for the summer. Those were times of sun, fun, and excitement when we could freely explore, learn, and understand the fascinating mysteries of life and living. We were comfortable in our secure little worlds with familiar people, traditions, and surroundings. As we basked in the sunlight of yesterday, we hoped that things would never change, but somehow we knew they would.

By analogy, we are facing the same challenge in the accounting profession, which has undergone two revolutionary events in its history. The first event occurred when Luca Pacioli published his famous book entitled, "The Collected Knowledge of Arithmetic, Geometry, Proportion, and Proportionality." The second event, digital revolution, just took place over the past few decades. Between the two events, accounting methods and procedures evolved and adapted to environmental changes as they occurred, which for the most part were relatively slow and predictable. Then suddenly, dramatically, and permanently, digital technology disrupted our peaceful existence. Ah, if we could linger and relax in the warm sunlight of our childhood summers for just a little while longer, perhaps the digital revolution will somehow miss us and we might continue as if nothing happened. As with children, we hope that things will never change, but deep inside we know they must.

The accounting profession has responded to the digital revolution quite admirably by continuing to refine and develop accounting standards, procedures, and services appropriate to the new global, technology-centric business world. Simultaneously, the profession should look to accounting researchers for further guidance in this regard, as the survival instinct of professionals (quite understandably) is to react to environmental changes while academics are supposed to take a complementary proactive stance that can help the profession survive and prosper in the long-run. Unfortunately, some accounting scholars refuse to adapt their research foci accordingly; rather, they remain fixated on stale questions and outdated approaches. Recently, some prominent accounting researchers have stated that scholars should not respond to needs of the profession; instead, we should continue to conduct "basic" research. While I agree that researchers must "stay the course" and be careful not to follow the profession through every treacherous bend in the road, we should not ignore systemic changes taking place in the world around us, as some academics would prefer. It is as if, like children, they hope to play in the same old sandbox until they retire, leaving those who remain behind to pick up their obsolete broken toys.

My contention is that the digitization of economic phenomena across the globe has fundamentally and permanently changed the accounting profession.

Accordingly, a large and growing proportion of "basic" research must incorporate the impact of information and communication technologies (ICT) on the psychology and behavior of accountants, as well as other consumers of accounting information. Admittedly, all research topics do not necessarily need to integrate ICT considerations; for instance, when investigating the reaction of investors to earnings pre-announcements, researchers need not worry about ICT, as the research question deals with decision-making processes that are independent of technology. However, an expanding set of behavioral research questions in accounting is inexorably intertwined with ICT. The following sections will explicate this thesis.

DIGITIZATION AND ACCOUNTING PRACTICE

The digitization of economic and accounting phenomena has dramatically and eternally altered the business landscape before us. We now stand on the brink of a precipice. As we look down into the abyss, we cannot see the bottom for the dark envelops and absorbs the light. We are unsure how to traverse the caverns, rivers, and canyons to get to the other side, as there are no marked trails; in fact, we cannot even see the other side. However, we know that we must not turn back. We can view the digital world before us with fear, apprehension, and resistance, or we can envision a new business and accounting panorama that is filled with exciting opportunities, challenges, and rewards.

Were one to measure how much business practice has changed over the past three decades, due in most part to digital technology, we could use fluxion mathematics, from which calculus arose, to conceptually calculate the distance traveled by dividing the "change in practice" by the "change in time". There is little doubt that such a "distance function" would reveal that business practice has traveled a very long way indeed. Next, if we could perform another conceptual calculation, the first derivative of the distance function would likely indicate that the "velocity" of change has been breathtaking. Most would agree that the speed with which change has taken place over the last 30 years has been faster than in the entire history of recorded mercantile exchange. A final conceptual calculation involving the second derivative of the distance function would most certainly reveal that "acceleration", or the rate of change in speed, has not slowed over the past three decades. In fact, we would likely find that the acceleration of change in business practice reflects an exponential upward slope with no end in sight.

The acceleration function just described suggests a crushing "G" force, which places immense pressure on business executives to stay ahead of the power curve by leveraging on ICT to dramatically and continually redesign business processes

and models to keep pace with onerous competitive demands. The accounting profession has not escaped the ubiquitous and disruptive, yet enabling power of ICT, as accountants have been forced to adapt their compilation, audit, tax, and managerial practices and procedures accordingly. Unfortunately, they have done so with little meaningful guidance from accounting researchers.

The manner in which humans respond to change tends to fall into three groups: some embrace change without a grudge, others wish to adapt but need a nudge, and a few absolutely refuse to budge. A small but growing number of accounting scholars are already weaving ICT considerations into their research questions; thus, to some people I am preaching to the choir (i.e. the first group). A stubborn and hopefully shrinking number of researchers remain steadfast in their belief that ICT represents a newfangled fad that will eventually go away (i.e. the third group); hence, they will refute all points made in a commentary of this nature, as they choose to remain locked in their little worlds. Accordingly, my remarks herein are primarily aimed at researchers who recognize the need to incorporate ICT factors into their scholarly pursuits, but seek guidance toward this end (i.e. the second group).

DIGITIZATION AND ACCOUNTING BEHAVIORAL RESEARCH

Since the entire spectrum of accounting research cannot be adequately dealt with in a brief commentary, the following discussion focuses on the potential impact of digitization on the nature and quality of accounting behavioral research. The "nature" aspect addresses the following question: "What has changed in accounting that requires researchers to incorporate ICT factors into their investigations?" The "quality" aspect deals with how scholars can incorporate ICT into their research designs to improve the internal validity and productivity of their projects.

DIGITIZATION AND RESEARCH RELEVANCE

The days of manually entering, authorizing, processing, monitoring, controlling, reconciling, and reporting transaction-based data for economic entities are quickly fading into the sunset. While some level of manual transaction processing will likely remain for highly uncertain and ambiguous economic events, most of the lower-level work previously performed by accountants is now or soon will be automated. Hence, the very nature of work is rapidly changing for accountants due to the infusion of ICT. For instance, beginning

with data capture through knowledge management, the way in which corporate accountants think about handling economic events in a digital business environment bears little resemblance to yesteryear. The same can be said for auditors, tax professionals, and consultants. The following discussion offers the digital accounting value chain (see Fig. 1) as a framework for pondering how accounting behavioral researchers can, and why they should, integrate ICT considerations into many of their research projects.

Early attempts at integrating organizations across horizontal business processes, rather than down vertical functional areas, often failed primarily due to the lack of a supporting ICT infrastructure. Hence, the concept of business process redesign fell out of favor after a while, as firms grew weary of expending huge resources on redesign efforts that yielded few tangible rewards. However, the capability, speed, and affordability of digital technology eventually matured to the point where organizations could begin to realize the full potential of business process redesign efforts. One of the most fruitful manifestations of technological innovation and maturation has been the advent of enterprise resource planning (ERP) systems, which allow for full integration of business processes across the organization. ERP systems can be used to digitally capture economic events, reliably process information, and rapidly disseminate knowledge. As a result, many accountants now handle fewer mundane tasks than in the past; instead, they are expected to add value to organizations by employing higher-level analytical and technical skills. To visualize how ICT has changed the nature of accounting work, assume that a manufacturer of laundry detergents, called Clean Manufacturing, has implemented a fully integrated ERP system.

Capturing Economic Phenomena

The integrated ERP system features on-line self-service applications that allow distributors to place product orders, retailers to submit shelf placement ideas, customers to express product quality concerns, and employees to maintain benefit selections. The distributors' orders are automatically translated into production requests and sent to the manufacturing function, the retailers' ideas are made readily available to marketing managers, the customers' quality issues are sent to key players in sales, production, and general management, and the employees' changes are transmitted to the third-party benefit provider.

Where are the accountants in this scenario? In past years, accountants were used primarily to code, enter, edit, and process input data of this nature. Now, most input is entered at the source and processed automatically. Certainly there are productivity and internal control issues to consider in this scenario, as prac-

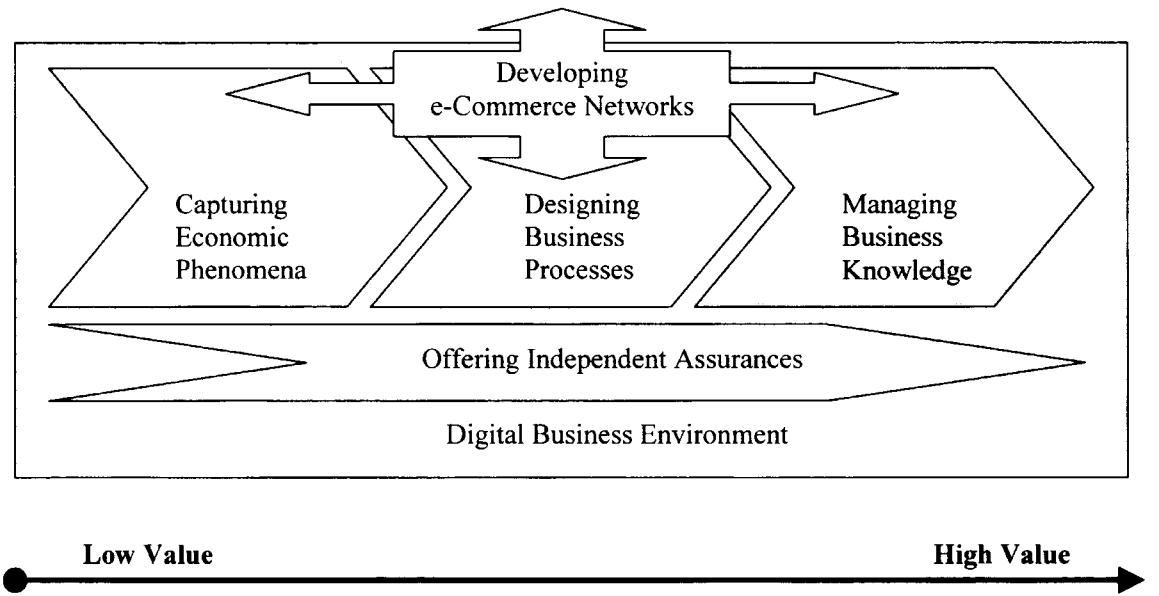


Fig. 1. The Digital Accounting Value Chain.

Adapted from "The Impact of Information and Communication Technology on Business Practices, Accounting Professionals, and Academic Researchers," by J. E. Hunton, *Accounting Horizons*, forthcoming.

ticing accountants can be useful in designing innovative digital input processes and ensuring the integrity of captured data via programmed input controls. Hence there are educational ramifications, such as teaching accounting students how advanced accounting information systems function, particularly with respect to internal controls.

Additionally, there are research opportunities for behavioral scientists on the input side of the equation. For instance, accounting behavioral researchers can study the most effective way to design input screens and event-capturing devices to optimize human-to-computer interactions. Research of this nature has deep roots in information systems and cognitive science, which has resulted in the development of several theoretical models such as the Technology Acceptance Model (TAM) (Davis, 1989, 1993; Davis et al., 1989), Task-Technology Fit (TTF) Model (Goodhue, 1995; Goodhue & Thompson, 1995), and the Media-Task-Fit (METAFIT) Model (Nöteberg et al., 2002). Also, in situations where security and privacy concerns abound, behavioral researchers can examine the types of input measures that will offer "risk relief" in this regard to affected parties. Hunton et al. (2000) offer an example of such research and much more needs to be investigated along this line of inquiry.

Designing Business Processes

Next, the ERP system schedules and coordinates workers, machines, and materials for upcoming production runs, based on distributors' orders. The manufacturing resources simultaneously arrive where and when needed, the goods are produced, and the products are shipped to the requested distributor locations. The ERP system then notifies the distributors of shipping dates, shipping terms, and expected arrival dates via automated e-mail confirmations. The feedback loop just described effectively integrates digital inputs, information processes, and business events into a real-time digital response network with little or no human intervention.

Once again, we find ourselves asking the question: "Where are the accountants in this scenario?" While their involvement may not be apparent on the surface, as most information processing is automatic, the real value of contemporary accountants begins to shine behind the scenes. Because accountants fully understand the flow, interrelationship, and control of business information throughout the entire organization, they can serve as catalysts for weaving innovation into business processes. This requires a great deal of creativity with respect to business process design issues and a high level of competency in the area of computer technology. After all, how can accountants effectively improve business processes if they are unaware of how to leverage ICT in this regard?

Certainly, there are educational issues here, but there are behavioral implications, as well.

For instance, behavioral researchers can examine effective ways to help accounting practitioners and students develop their innate creativity. Toward that end, behavioralists can compare various schematic representations of "reality" to determine which model conveys the most understanding under what circumstances. Research of this nature has already begun with respect to semantically modeled accounting systems (e.g. Gerard, 1999; Dunn & Grabski, 2000; O'Leary, 1999). Additionally, during and after business process redesign, researchers can study the best means of facilitating change in an organizational setting. Some accounting researchers are already examining such issues under the umbrella of "participation" research (e.g. Essex, 1997; Whitecotton & Butler, 1998; Hunton & Gibson, 1999). Last, but certainly not least, a handful of accounting researchers are investigating the ethical implications of displacing workers with automated processes, as there are personal and organizational ramifications in doing so inappropriately (e.g. Arnold et al., 1997; Yuthas & Dillard, 1998; Sutton et al., 1999).

Managing Business Knowledge

Next, the integrated system continuously scans and reports meaningful information to company managers, who use their experience and judgment to transform such information into valuable decision-making knowledge. For instance, marketing managers at Clean Manufacturing might heed the advice of retailers and move some products from middle shelves to end-cap locations. Research and development managers might investigate customer complaints and, as a result, alter the formula of certain products to enhance their cleaning power. Production managers might identify upcoming machine scheduling conflicts and modify resource allocations accordingly. Decision scenarios of this nature can be incorporated into the organization's knowledge base, which can be used to identify and enhance "best practices" in the future.

We are now on the high-value end of the accounting value chain, where practising accountants and academic researchers can add considerable value. For instance, researchers can investigate the extent to which more comprehensive reporting of financial and non-financial events, such as the balanced scorecard, facilitates better understanding, comprehension, and decision-making by various stakeholder groups, as recently examined by Lipe and Salterio (2000). As another example, researchers could examine the market effect (e.g. stock price volatility) of continuous financial reporting made possible by ICT and XBRL. While economic theory suggests that continuous reporting will smooth market

volatility, psychological theory (e.g. decision heuristics and biases) suggests that volatility will increase. Additionally, questions regarding the extent to which continuous reporting might affect quality of earnings should be addressed; for example, if continuous reporting constrains management's ability to switch accounting estimates, accruals, disclosures, and choices throughout the year without raising earnings management "red flags", will quality of earnings improve? To date, no research has been published with respect to the two issues just mentioned. Other research opportunities abound in the areas of decision aids (Johnson & Kaplan, 1996; Boatsman et al., 1997), group support systems (e.g. Bamber et al., 1996; Arnold et al., 2000) and expert systems (e.g. Murphy & Brown, 1992; Sangster, 1996), as such systems can leverage on internally and externally developed knowledge to improve human decision-making. Finally, there are many behavioral research issues surrounding ways to capture, process, and disseminate knowledge so that a wide range of decision-makers can convert information to knowledge, knowledge to decisions, and decisions to wisdom (e.g. Bartlett, 1998; O'Leary, 2000).

Developing e-Commerce Networks

Assume that Clean Manufacturing receives immediate digital inputs from retail stores that allow managers to monitor inventory levels, observe product locations, and measure stock flows at retail stores. Real-time intelligence gathering of this nature would allow Clean Manufacturing to offer valuable guidance to e-commerce partners, such as alerting distributors where and when to stock shelves and suggesting product placement strategies to retailers.

A digital e-commerce network also facilitates the emergence of new business models; for instance, manufacturers could assume total responsibility for their products from production to consumption. A shift of this nature would give manufacturers greater control over their products, while relieving financial and human resource burdens heretofore imposed on distributors and retailers. Taken one step further, manufacturers might question why they use distributors at all, since they can now automate and manage most distributor functions. Another possible business model is one in which manufacturers operate their own Internet stores for direct sales to customers, thereby bypassing retailers entirely.

Additionally, the digital e-commerce network allows Clean Manufacturing to track the stock of raw materials held by primary and secondary suppliers, view production schedules of alternate manufacturers with whom it has reciprocal contingency production agreements, and obtain inventory levels of products at distributor locations (relayed via shelf sensors). At some point, based on digitally

captured input and process information, the ERP system at Clean Manufacturing can automatically generate production orders. As noted earlier, workers, machines, and materials are then scheduled to arrive where and when needed. Because the ERP system is linked to the suppliers' internal information systems, Clean Manufacturing can trigger purchase orders when needed and, upon receipt of materials, electronically transfer payments to suppliers. You guessed it – this all occurs with little or no human intervention.

On the output side of the e-commerce equation, decision-makers throughout the organization can constantly monitor the digital network, modify system parameters on the fly, and reallocate resources in response to changing conditions. For example, sales of Product B might be rapidly slipping in Australia while suddenly rising in Germany. The company can divert necessary resources associated with Product B to Germany in an effort to keep up with increasing demand, while simultaneously attempting to understand reasons behind the declining sales in Australia. After thorough analysis, management's reaction might be to pour more marketing resources into Australia, investigate successful sales tactics in Germany, solicit customer feedback from both countries, and so on. The important point here is that the digital network provides the technological infrastructure to deliver reliable real-time information and knowledge up, down, and across the digital accounting value chain.

The behavioral research issues associated with external linkages are much the same as previously mentioned in conjunction with capturing economic inputs, processing business information, and managing business knowledge. However, the scope expands because digital networks allow multiple parties with various, and sometimes conflicting, motives to engage in e-commerce relationships in ways heretofore not possible without a sophisticated ICT infrastructure. Hence, existing and emerging e-commerce models demand that accountants take a holistic look at the entire business environment from marketing to production to distribution, particularly with respect to the conveyance of more timely and relevant business information for decision-makers within and outside the network.

Offering Independent Assurance

The most obvious manifestation of independent assurance is found with the auditor's opinion on the fairness of financial reporting. While this line of assurance is likely to continue into the foreseeable future, the nature, timing, and extent of such assurance is likely to change dramatically. For instance, coupled with continuous financial reporting is the related issue of continuous assurance. Will the capital market ascribe added value to continuous assurance? Will

investors place more value on one type of information (e.g. financial results) over another type (e.g. non-financial announcements), and if so, how, why, and under what circumstances? How will company managers react to continuous assurance? Will they change their decision models out of fear (or perhaps retribution) of perceived "big brother" oversight and snooping? Will quality of earnings be impacted one way or the other by continuous assurance? Another line of inquiry with regard to financial auditing concerns unique risks posed by advanced information systems. For example, are there risk differences between a legacy system and ERP system? What is the nature of such risk difference and what is the risk exposure? Do traditional financial auditors understand unique business and audit risks posed by ERP systems? These and many more behavioral research questions arise in the context of financial reporting assurance.

Additionally, the interjection of ICT infrastructures into the business world opens the door to many new assurance services. One of the most promising assurance services is SysTrust^{SM}, which has business-to-business orientation. With SysTrust^{SM}, auditors offer assurances with respect to the reliability (i.e. availability, security, integrity, and maintainability) of a client's information system, based on the client's assertions. A SysTrust^{SM} report issued by independent auditors can be used for many purposes, such as lowering a firm's cost of capital or providing risk relief to potential trading partners. Several behavioral issues immediately come to mind in this regard. For instance, if an application service provider obtains a SysTrust^{SM} report, does it really matter to potential service recipients; meaning, does systems reliability assurance reduce perceived transaction risk? If so, which aspect(s) of systems reliability matter most to potential service recipients? Does the CPA brand offer enough elasticity to stretch into assurance areas other than the traditional audit such that CPA's are deemed to be competent in these areas? If not, what actions should be taken to facilitate brand transference of this nature? Similar behavioral issues arise in the context of each new assurance service offering.

Finally, advanced ICT infrastructures allow auditors to work in near virtual reality, where space and time become less important. For instance, audit work teams across the globe can share information related to the same client via workgroup software and group support systems. Many behavioral issues arise in this context, such as how to design effective yet efficient electronic work papers, how to best assess collective risk and materiality for a geographically disbursed client, how to develop rich media with a high degree of social presence so that auditor-to-auditor and auditor-to-client relationships are properly cultivated, and how to improve audit judgment in a technology-centric business environment.

DIGITIZATION AND RESEARCH QUALITY

ICT has also changed the way in which research can be conducted. Without delving into too many issues here, I will use the context of two studies in which I was involved to illustrate how ICT can be used to improve the quality of accounting behavioral research.

Hunton and McEwen (1997) used computer technology to track the eye movements of financial analysts to ascertain their cognitive search strategies while they gleaned information from traditional financial reports and made earnings forecasts. Prior studies of financial statement users employed verbal protocol analysis to determine the kind of financial information sought and the search pattern employed. While verbal protocol analysis is useful in this regard, some researchers have expressed concern that the obtrusiveness of the procedure can alter search patterns and cognitive processes. Accordingly, we used a less obtrusive method via retinal imaging technology to examine similar issues.

During the experiment, we collected volumes of data concerning what the analysts saw, how long they spent on each piece of information, and in what order they proceeded through the financial statement items (including footnote disclosures). As a result, we could mathematically determine whether they employed a directive or sequential search pattern and link their search strategy to their experience level, among other factors. Additionally, computer technology allowed us to completely randomize (using a random number generator) participants to treatment conditions, the order in which each participant saw financial statement items, the order of dependent variable questions, and the order of manipulation check items. Finally, we were allowed to place the computerized experiment on-site at a national stock brokerage firm where volunteers could participate on their time schedule. This is but one example of how technology can be used to improve the quality of behavioral research methods, particularly with respect to improved internal validity.

Beeler and Hunton (2001) offer another example of how ICT can be used to improve research quality, and in this case, productivity. The objective of this study was to determine the extent to which audit judgment can be biased in light of contingent economic rents, which is at the heart of the independence debate. To examine this issue, we created case materials from an actual company that went bankrupt. We allowed participants to review historical five-year financial statements and ratios of the company prior to bankruptcy, and then asked them to rate the likelihood that the company will remain viable in the coming year (i.e. going concern judgment). The experiment employed a two (low balling: absent or present) by two (potential non-audit revenues: absent or present) design.

As with the prior study, we randomized treatments to participants, and the order of informational items, dependent variables, and manipulation checks. The next step was to find the most appropriate participants given the nature of the study. To accomplish this, through our networks of professional acquaintances in the auditing profession, we solicited the participation of audit partners from four of the Big Five CPA firms. Realizing that audit partner time is valuable, we wanted to make the experiment available over the Internet so that the partners could complete the experiment when and where they desired. We built several computerized controls into the experiment to ensure the proper identity of the volunteer audit partners and placed the experiment on a web server. In the end, 73 audit partners across the nation participated in the experiment within a five-day window of time. Thus, not only was the internal validity of the study improved, but research productivity was accelerated too.

SUMMARY

One objective of this commentary is to stir the imagination of behavioral researchers with respect to emerging research opportunities brought about by the infusion of advanced information and communication technologies (ICT) into business and accounting environments. The second objective is to offer insight with regard to how ICT can be used to strengthen the internal validity of behavioral experiments, as well as research productivity. While a short commentary of this nature cannot possibly cover all research ramifications of ICT, perhaps the mention of a few ideas will inspire some accounting scholars to incorporate innovation and creativity into their research projects in this regard. In doing so, we can hopefully lead the accounting profession through perilous waters of the digital revolution via our research endeavors, rather than follow behind extant practice.

REFERENCES

Arnold, T., Arnold, V., & Sutton, S. G. (1997). Toward a philosophical foundation for ethical development of audit expert systems: A contractarian approach. *Research on Accounting Ethics, 3*, 211–232.

Arnold, V., Sutton, S. G., Hayne, S. C., & Smith, C. A. P. (2000). Group decision making: The impact of opportunity-cost time pressure and group support systems. *Behavioral Research in Accounting, 12*, 69–96.

Bamber, E. M., Watson, R., & Hill, M. (1996). The effects of group support systems on audit group decision making. *Auditing: A Journal of Practice and Theory, 15*(1), 122–134.

Bartlett, C. (1998, April 20). Managing knowledge and learning. *Harvard Business School, 9*, 396–357.

Beeler, J. D., & Hunton, J. E. (2001). Contingent economic rents: Insidious threats to auditor's judgments. *Advances in Accounting Behavioral Research, 5*, forthcoming.

Boatsman, J. R., Moeckel, C., & Pei, B. K. (1997). The effects of decision consequences on auditors' reliance on decision aids in audit planning. *Organizational Behavior and Human Decision Processes, 71*, 211–247.

Davis, F. D. (1989, September). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319–339.

Davis, F. D. (1993). User acceptance of information technology: System characteristics, user perceptions and behavioral impacts. *International Journal of Man-Machine Studies, 38*, 457–487.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*, 982–1003.

Dunn, C. L., & Grabski, S. V. (2000). Perceived semantic expressiveness of accounting systems and task accuracy effects. *International Journal of Accounting Information Systems*, forthcoming.

Essex, P. (1997). Resistance to system development: An equity-based model. *Advances in Accounting Information Systems, 5*, 167–204.

Gerard, G. (1999). Experience effects and the role of knowledge structure in accounting systems design. Unpublished Ph.D. dissertation, Michigan State University.

Goodhue, D. L. (1995). Understanding user evaluations of information systems. *Management Science, 41*, 1827–1844.

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly, 19*, 213–237.

Hunton, J. E., Benford, T., Arnold, V., & Sutton, S. (2000). The impact of electronic commerce assurance on financial analysts' earnings forecasts and stock price estimates. *Auditing: A Journal of Practice and Theory, 19*, 5–22.

Hunton, J. E., & Gibson, D. (1999). Soliciting user-input during the development of an accounting information system: Investigating the efficacy of group discussion. *Accounting, Organizations and Society, 24*, 597–618.

Hunton, J. E., & McEwen, R. A. (1997). An assessment of the relation between analysts' earnings' forecasts, motivational incentives, and cognitive information search strategy. *The Accounting Review, 72*(4), 497–516.

Johnson, V. E., & Kaplan, S. E. (1996). Auditor's decision-aided probability assessments: An analysis of the effects of list length and response format. *Journal of Information Systems, 10*(2), 87–102.

Lipe, M. G., & Salterio, S. E. (2000). The balance scorecard: Judgmental effects of common and unique performance measures. *The Accounting Review, 75*(3), 283–298.

Murphy, D. S., & Brown, C. E. (1992). The uses of advanced information technology in audit planning. *International Journal of Intelligent Systems in Accounting, Finance & Management, 1*(3), 187–193.

Nöteberg, A., Hunton, J. E. & Benford, T. (2002). Optimizing the 'fit' between electronic communication media and tasks. *International Journal of Accounting Information Systems*, forthcoming.

O'Leary, D. E. (1999). REAL-D: A schema for data warehouses. *Journal of Information Systems, 13*(1), 49–62.

O'Leary, D. E. (2000). Management of re-engineering knowledge: AI-based approaches. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 9(2).

Sangster, A. (1996, November). Expert system diffusion among management accountants. *Journal of Management Accounting Research, 8*, 171–182.

Sutton, S. G., Arnold, T. D., & Arnold, V. (1999). An integrative framework for analysis of the ethical issues surrounding information technology integration by the audit profession. *Research on Accounting Ethics*, *5*, 21–36.

Whitecotton, S. M., & Butler, S. A. (1998, Supplement). Influencing decision aid reliance through involvement in information choice. *Behavioral Research in Accounting*, *10*, 182–200.

Yuthas, K., & Dillard, J. (1998). Structuration theory: A framework for behavioral accounting research. *Advances in Accounting Behavioral Research*, 195–221.

CONTINGENT ECONOMIC RENTS: INSIDIOUS THREATS TO AUDIT INDEPENDENCE

Jesse D. Beeler and James E. Hunton

ABSTRACT

The primary purpose of this study is to examine how and why contingent economic rents can lead to biased audit judgment via a cognitive processing phenomenon known as predecisional distortion of information. The secondary research objective is to develop a more refined measure of predecisional distortion, thereby yielding a more robust and predictive metric. A total of 73 audit partners representing four of the Big Five CPA firms participated in a two (low-balling: present or absent) by two (non-audit revenue: present or absent) between-subjects experiment. Research findings indicate that contingent economic rents can potentially impair audit independence, as such rents heighten an initial desire to maintain a long-term relationship with the client, trigger favorable predecisional distortion of client-related information, and bias audit judgment in favor of the client. Study results also reveal that the refined predecisional distortion metric is more predictive than past measurement techniques. Between-subject and within-subject debriefings suggest that predecisional distortion of information operates, at least partially, at the subconscious level.

Advances in Accounting Behavioral Research, Volume 5, pages 21–50.

ISBN: 0-7623-0953-9

INTRODUCTION

Recently, the Securities and Exchange Commission (SEC) revamped the auditor independence rule. Among other things, the revised rule limited the nature and extent of non-audit services that Certified Public Accountant (CPA) firms can provide to registered audit clients (SEC, 2000). The SEC's primary motivation for re-examining the independence rule arose over its concern that contingent economic rents (i.e. future earnings from audit clients that are conditioned on maintaining on-going audit relationships with the clients) might bias auditors' judgments in favor of their client, thereby potentially impairing independence.

Prior auditing research suggests that the SEC's anxiety in this regard is not totally unfounded, as some studies have reported that auditors who hold an initial hypothesis or outcome preference tend to engage in confirmatory processing while gathering and evaluating client-related information (e.g. Bazerman et al., 1997; Church, 1990; Smith & Kida, 1991). Extant psychology research also indicates that it is difficult for decision-makers to remain unbiased, objective, and impartial during information evaluation processes when they hold an initial desire for a given outcome (Chapman & Elstein, 2000; Kunda, 1990; Russo et al., 1996, 1998, 2000).

The primary purpose of this study is to investigate how and why contingent economic rents can bias the professional judgment of auditors and potentially impair audit independence. In particular, we focus on a cognitive process phenomenon called predecisional distortion of information, which refers to the amount of bias incorporated into the information evaluation phase of decision-making before arriving at a final judgment or choice. The secondary objective of the current research is to refine the manner in which predecisional distortion has been measured in past studies (e.g. Russo et al., 1996, 1998, 2000), thereby yielding a more robust and predictive metric.

In the current experiment, 73 audit partners representing four of the Big Five CPA firms were asked to assume that they had just retained an audit client described in case materials. Two forms of contingent economic rents were manipulated, resulting in a two (low-balling: present or absent) by two (non-audit revenue: present or absent) between-subjects experimental design. As hypothesized, the presence of contingent economic rents elevated the auditors' initial desire to maintain a long-term relationship with their client, triggered favorable predecisional distortion of client-related information, positively affected preliminary likelihood assessments that the company would continue as a going concern, and negatively influenced initial budget hour revisions. Study results further indicate that the refined predecisional distortion metric is more predictive than past measurement techniques. Results of between-subject

and within-subject debriefings reveal that predecisional distortion and biased audit judgment appear to operate, at least partially, at the subconscious level.

This study contributes to extant auditing and psychology literature in several key areas. First, two forms of contingent economic rents are identified as exogenous factors that activate predecisional distortion of information during the evaluation of audit evidence. Second, the psychological process through which predecisional distortion can bias audit judgment is articulated. Third, the concept of predecisional distortion is expanded to include decision weights, thereby yielding a more refined and predictive distortion index. Finally, research evidence suggests that confirmatory processing and biased judgment are not necessarily volitional; rather, psychological factors operating at the subconscious level appear to influence auditors in this regard.

Section II advances the theoretical framework for this study and proposes research hypotheses. Section III describes the research method, Section IV presents the experimental results, and Section V discusses the research findings.

THEORY AND HYPOTHESES

For the most part, prior definitions of independence assume that violations occur when auditors knowingly misrepresent the truth (e.g. DeAngelo, 1981; Antel, 1984; Simunic, 1984; Magee & Tseng, 1990; Lee & Gu, 1998). While deliberate misrepresentation may occur in rare instances, we suggest that independence impairment is more often the result of subconscious biases that alter and distort the collection and evaluation of audit evidence, such that auditors seek and interpret evidence in a manner consistent with their desired outcomes. The process of gathering and assessing audit evidence in this fashion is known as confirmation bias (Nisbett & Ross, 1980) or confirmatory processing (e.g. Fischhoff & Beyth-Marom, 1983; Church, 1990).

As depicted in the research model (see Fig. 1), the current study proposes that when the realization of future economic rents depends on maintaining ongoing audit relationships with clients (the contingency), auditors are favorably predisposed toward their clients (the initial preference); thus, predecisional distortion is activated during the evaluation of audit evidence. In turn, higher levels of pro-client bias reflected in predecisional distortion are expected to positively influence going concern likelihoods and negatively affect budget hour revisions. Because confirmatory processing serves to align cognition (audit evidence), motivation (desire to keep the audit client) and outcomes (audit judgment), we suggest that auditors will believe they are rendering impartial, objective, and unbiased judgments, but will be making a decision that is biased in favor of the audit client.

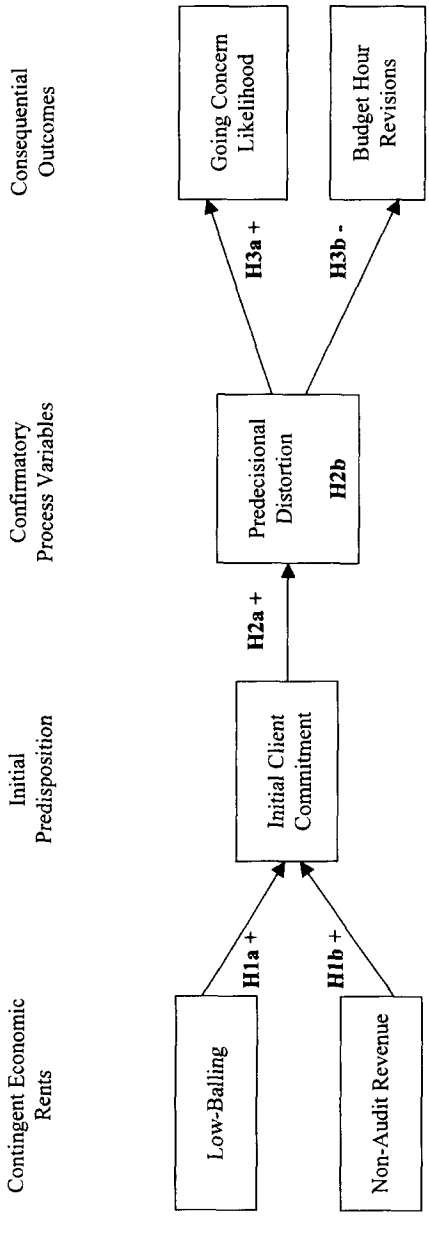


Fig. 1. Research Model.

While not depicted on the research model, we also recognize the likely existence of significant direct paths from the independent to dependent variables, particularly since contingent economic rents are experimentally manipulated and going concern assessments and audit plans involve a host of endogenous and exogenous factors not considered in this study. The following sections discuss each construct and relationship presented in the research model.

Low-balling

Low-balling is a price cutting strategy whereby the initial year engagement fee is less than subsequent year fees (Baber et al., 1987; Francis & Simon, 1987; Simon & Francis, 1988; Schatzberg, 1990; Lee & Gu, 1998). One purpose of employing such a pricing strategy is to gain access to potential client-specific economic rents in future periods in the form of higher audit fees (Farmer et al., 1987).

The existence of low-balling strategies and impact of low-balling on auditors has been investigated to some extent in prior research. For instance, in an archival study, Simon and Francis (1988) observe a 25% reduction in initial engagement fees upon a change of auditors, as compared to normal levels of on-going audit fees. Audit fees in the second and third years remain 15% below normal levels and then climb to normal levels by the fourth year. In an experimental market simulation, Schatzberg (1990) conclude that low-balling is most likely to occur in competitive markets where transaction costs are positive. Several studies have examined the influence of low-balling on auditor's judgments (e.g. Schatzberg & Sevcik, 1994; Schatzberg & Shapiro, 1996; and Calegari et al., 1998). For the most part, research in this area suggests that auditor judgment is potentially biased in favor of the client in the presence of a low-balling strategy. While the collection of future audit fees reflects one form of economic rents, the provision of non-audit services represents another form.

Potential Non-Audit Revenue

For decades, the SEC has questioned whether auditors should provide non-audit services to audit clients (SEC, 2000). Empirical evidence on the extent to which audit quality is compromised by contingent economic rents arising from potential non-audit revenue is limited and equivocal. For instance, Titard (1971) and Pany and Reckers (1984) report that financial statement users are highly concerned about independence impairment when auditors provide consulting services to audit clients, but are less concerned when a separate division of the

CPA firm offers such services. However, Schockley (1981) finds that financial statement users are concerned about auditor independence in both scenarios just described. Using analytical modeling, Simunic (1984, p. 699) notes, "The involvement of the auditor in MAS [management advisory services] reduces the probability of truthful audit reporting if the MAS work generates economic rents." A survey of loan officers and financial analysts indicates that auditor provided consulting services do not impact their perceptions of auditor independence (McKinley et al., 1985; Pany & Reckers, 1988). Yet, another study of loan officers suggests that the provision of consulting services would likely result in audit conflicts being resolved in favor of the client (Knapp, 1985). In an experimental markets study, Schatzberg and Shapiro (1996) indicate that potential economic rents in the form of non-audit fees can impair auditor independence if the benefits of misrepresenting audit findings exceed the costs.

While prior studies in the area of low-balling and non-audit revenues suggest that audit judgment can be biased in the presence of contingent economic rents, the psychological process through which such bias is created remains unclear. We suggest that contingent economic rents heighten auditors' commitment to retain their client (the contingency and initial anchor), which leads to predecisional distortion of client-related information and biased audit judgment, as next discussed.

Initial Predisposition

The motivated reasoning literature (e.g. Pyszcaynski & Greenberg, 1987; Kunda 1990) suggests that the economic self-interests of audit partners, arising from contingent economic rents, will establish an initial pre-disposition toward retaining their audit clients. An anchoring phenomenon of this nature is activated by a self-serving bias (Messick & Sentis, 1979), which means that an individual's desired outcome is heavily weighted toward the alternative that best serves his/her self-interest. The self-serving-bias is difficult to suppress (Jarvis & Petty, 1996) and represents a powerful, yet subconscious, force that impacts the manner in which auditors subsequently process audit evidence (Bazerman et al., 1997; Russo et al., 2000). Accordingly, we propose that the presence of contingent economic rents will heighten auditors' commitment toward retaining their client, as indicated by the following hypothesis:

H1a: There will be a positive relationship between low-balling and initial client commitment.

H1b: There will be a positive relationship between non-audit revenues and initial client commitment.

Predecisional Distortion of Information

Once a preferential outcome is established (e.g. client retention), auditors will exhibit a tendency to evaluate client-related evidence in a manner consistent with their initial preference (e.g. Lord et al., 1979; Gorman, 1986; Church, 1990; Smith & Kida, 1991; Russo et al., 1996; Bamber et al., 1997; Lweicka, 1998; Nickerson, 1998; Chapman & Elstein, 2000; Wilks, 2001). Evaluating audit evidence in this manner is known as confirmatory processing. There has been a great deal of research in this area (e.g. Kida, 1984; Libby, 1985; Church & Schneider, 1993; Bedard & Biggs, 1991; Brown et al., 1997; Cloyd and Spilker, 1999; Church, 1991; Johnson, 1993; McMillian & White, 1993). For the most part, research evidence indicates that auditors are prone to a confirmatory processing bias.

One means of aligning audit evidence with initial preferences is known as predecisional distortion of information (e.g. Mongtomery, 1983; Sevenson, 1992; Shafir, 1993). On the whole, decision-makers tend to distort information in favor of an initially preferred outcome or hypothesis (e.g. Elliot & Devine, 1994, Frey, 1986, Smith & Kida, 1991). To test the extent to which predecisional distortion of information affects preferential choices, Russo et al. (1996, 1998) provided student participants with information regarding various attributes of two alternatives. Their research findings revealed that initial preferences for a given alternative significantly influenced the participants' favorability ratings of informational (attribute) items, such that the preferred alternative received higher attribute ratings. Using another series of binary preferential choice tasks, Russo et al. (2000) found that predecisional distortion of information was pervasive and persistent for salespersons and auditors; however, the auditors' distortion was significantly less than the salespersons' distortion. A recent study by Wilks (2001) suggests that audit seniors exhibited predecisional distortion of audit evidence toward their supervisors' early views of the situation. Further, Wilks (2001) reports that, although auditors understood the biasing influence of supervisors' initial preferences on subordinates' judgments, they nevertheless exhibited predecisional distortion of client-related information.

Russo et al. (1996, p. 107) state, "Our findings might be seen as an addition to the literature on the confirmation bias in which a current belief or a desired conclusion leads to: (a) a search for confirming rather than opposing evidence; or (b) an overly confirming interpretation of the available evidence." Based on the theory and research presented above, we posit that heightened client commitment will lead to predecisional distortion of client-related information. Accordingly, the following hypothesis (alternate form) is offered:

H2a: There will be a positive relationship between initial client commitment and predecisional distortion of client-related information.

Refining the Measurement of Predecisional Distortion

In the studies cited above (Russo, 1996, 1998, 2000; Wilks, 2001), predecisional distortion was assessed via the participants' attribute ratings (e.g. favorable-unfavorable or optimistic-pessimistic) of each piece of information with respect to a choice preference or judgment task. However, the decision weight that participants assigned to each piece of information was not included in the measurement of predecisional distortion. We suggest that this procedure is potentially problematic, as subjective ratings for attributes that are relatively important and unimportant to the decision at hand are assumed to be equally weighted. As a result, a situation involving large rating changes (from control to treatment groups) for relatively unimportant attributes and small changes for relatively important attributes would represent a significant predecisional distortion effect. However, the extent to which such distortion impacts choice or judgment tasks might be minimal in this instance. In the previously cited studies, this did not appear to adversely impact the results, as the attributes may have been somewhat equally weighted; however, we suggest that a more robust measure of predecisional distortion will negate this potential threat to future studies. The notion of combining attribute ratings and importance ratings to form attitudinal and belief indices has been examined in psychology research.

For decades, researchers have investigated various phenomena related to multiple criteria decision-making (e.g. Fishbein, 1983; Keeney & Raiffa, 1976; VonNeumann & Morgenstern, 1947). Regarding subjective evaluations of object attributes, both attribute ratings (e.g. good/bad, favorable/unfavorable) and importance weightings (e.g. important/unimportant, high/low) are commonly measured. Multiplying the ratings and weightings assigned to each attribute, and summing the resulting product over all attributes, creates an attitude or belief index (Fishbein, 1983). The index is then linked to individual or group decisions in a variety of preferential choice and judgment tasks (Zeleny, 1984). In a commensurate fashion, assessing both attribute ratings and importance weightings, and then summing the product of ratings and weightings over all attributes, should yield a more refined and predictive predecisional distortion index.

Linking this concept to the current study, we suggest that, as compared to reflecting predecisional distortion by using either attribute ratings or importance weightings alone, the predecisional distortion index described above will best fit the research model. Hence, the following hypothesis is offered (alternate form):

H2b: A predecisional distortion index comprised of the multiplication of attribute ratings by importance weightings summed over all attributes will provide a statistically better model fit than either attribute ratings or importance weightings alone.

Consequential Outcomes

The case materials provided in the current study indicate a situation whereby the deteriorating nature of financial information and key ratios suggests a possible going concern problem. The case is based on an actual company that went bankrupt in the year following the audit. Accordingly, for one dependent variable, auditor partners are asked for their initial assessment regarding the likelihood that the client will continue to exist as a going concern in the coming year.[1] Confirmation bias theory suggests a positive relationship between confirmatory processing and going concern likelihood assessments, since predecisional distortion of information will place the client's financial position in a more positive light.

As a second dependent measure, participants are given an opportunity to revise initial budget hours. Given the uncertainty surrounding the client's financial condition, unbiased auditors will likely increase budget hours, particularly in the areas of analytical review and substantive testing, after learning about the client's precarious financial condition.[2] Accordingly, this study suggests a negative relationship between predecisional distortion and initial budget hour revisions, since the client's financial condition is expected to appear to be more positive to participants who exhibit higher levels of distortion. Therefore, the following relationships are hypothesized:

H3a: Predecisional distortion of client-related information will be positively associated with the auditors' initial likelihood assessments that the client company will continue as going concern in the coming year.

H3b: Predecisional distortion of client-related information will be negatively associated with revisions made to initial budget hours.

METHOD

The experiment employed a two (low-balling: present and absent) by two (potential non-audit revenue: present and absent) between-subjects randomized design. All subjects were provided with a computerized audit case. After an initial introduction screen, the participants read that they were the partners in charge of a new audit client. The absence [presence] of low-balling was next manipulated, as follows:

You are the partner who recruited this client in an effort to build the local office's audit practice. You bid the initial audit at above what it will cost to perform the audit; thus, you expect to make a profit on this engagement in the first year and you anticipate making profits on this engagement in subsequent years.

[You are the partner who recruited this client in an effort to build the local office's audit practice. You bid the initial audit at below what it will cost to perform the audit; thus, you expect to take a loss on this engagement in the first year, but you anticipate making profits on this engagement in subsequent years.]

The absence [presence] of potential non-audit revenue is manipulated on the same screen, as follows:

After holding preliminary discussions with the company's management team, you do not expect that your CPA firm will earn non-audit revenues (from management advisory services, tax planning, assurance services, information technology risk management, etc.) from this client in the future.

[After holding preliminary discussions with the company's management team, you believe that your CPA firm might earn non-audit revenues (from management advisory services, tax planning, assurance services, information technology risk management, etc.) from this client in the future.]

Next, the auditors' initial predisposition toward their client was assessed via the following statement:

At this point, I feel very committed to maintaining a long-term relationship with this client (1 = Strongly Disagree, 7 = Strongly Agree).

Participants are then asked to examine 12 pieces of client-related information by clicking on each of 12 buttons presented on a computer screen. The order of the buttons was randomized per participant to preclude an order effect. Participants read the following 12 categories of information: background, marketing, inventory system, net income & EPS, analyst reports, IPO information, income statements, balance sheets, liquidity ratios, solvency ratios, activity ratios, and profitability ratios. After reading all client information, the auditors are asked to rate the favorability and weight the importance of each piece of information, or attribute.[3] Auditors could re-examine each piece of information while responding to the following two questions for each informational item:

How would you characterize the information you just read with respect to assessing the financial viability of your client in the coming year? (–3 = Very Unfavorable, –2 = Somewhat Unfavorable, –1 = Slightly Unfavorable, 0 = Neutral, +1 = Slightly Favorable, +2 = Somewhat Favorable, +3 = Very Favorable).

How much decision weight do you place on the information you just read with respect to assessing the financial viability of your client in the coming year? (1 = Very Low Weight, 2 = Somewhat Low Weight, 3 = Slightly Low Weight, 4 = Neutral Weight, 5 = Slightly High Weight, 6 = Somewhat High Weight, 7 = Very High Weight).

A predecisional distortion index was calculated for each participant by multiplying the attribute rating by the importance weighting ascribed to each piece of information, and then summing the products across all 12 pieces of client information.[4]

Consequence Variable Measures

Next, participants provided their initial going concern likelihood assessments (0 to 100, in 10 point increments). Then, the audit partners read that the initial budget hours are 200 for analytical review, 500 for substantive testing, and 150 for audit administration, for a total of 850 hours (see note 2). Next, they were given an opportunity to revise the budgeted audit hours. The total hour change to the audit budget serves as the "revisions to initial budget hours" metric.

Other Measures

At this point, the computerized experiment prevented the audit partners from going back and changing prior responses. Participants next responded to a series of manipulation checks, between subject debriefing, and demographic information items which were randomized to preclude an order effect. Finally, at the end of the experiment, a within-subject debriefing was conducted to gain insight into whether the participating audit partners are aware that the presence of contingent economic rents could bias their own professional judgment.

Administration of the Experiment

The experimental materials were computerized and placed on a host computer so that the audit partners could participate in the study via the Internet. The partners were assigned individual passwords ahead of time, which authorized them to participate in the study. The participants were told that they must complete the entire study in a single session that would last about one hour.

Controls were built into the program such that each password granted a one-time authorization to the program. Once granted access, the participants were automatically randomized into treatment conditions. Oral interviews (mostly telephone) were conducted with all partners while they were in their offices. All partners indicated that they would participate in the study using their office computer. During the interviews, the researchers instructed the partners how to find the "computer name" listed in their network settings. As a further control to ensure that the partners were the actual participants, the host computer granted access to the experimental software only if the respondents

had the correct password and if the computer machine name matched their office computer.

All partners completed the study during a contiguous five-day window (Monday through Friday) wherein the experiment was available on the host computer. Several pilot tests of the experimental materials and computerized programs were conducted prior to the final experiment, using master's level accounting students and audit managers/partners from local and regional CPA firms. Necessary case materials and software changes were incorporated along the way.

RESULTS

Sample

The participants represent 73 audit partners from four of the Big Five accounting firms. The researchers initially approached 23 audit partners through their personal networks of professional contacts. The partners were told that the researchers were interested in examining how auditors arrive at client-related judgments. After agreeing to participate in the experiment, the partners then helped the researchers to recruit other partners within their firms to participate. A total of 75 participants were identified.

The researchers contacted all participants and obtained written voluntary consents. Complete anonymity and confidentiality were assured verbally and in writing to the participating audit partners. As an incentive to participate in the study, the researchers contributed $50 for each participant to the charity of their choice. Two of the 75 audit partners did not complete the experiment; accordingly, they were dropped from the study. The researchers personally called all participants after the experiment and debriefed them as to the study's purpose and summary results. Due to the sensitive nature of the study, after the debriefing, the partners were asked if they still wanted to have their data included in the study and all agreed in the affirmative. Sample statistics are presented in Table 1.

Results of statistical testing (X^2) indicate no significant proportional differences (all p-values exceeded 0.20) across treatment conditions for sample size, CPA firms, gender, or day of week (the experiment ran from Monday through Friday). Also, ANOVA testing revealed no significant differences (all p-values exceeded 0.50) across treatment conditions for any of the demographic variables. The overall mean (standard deviation) elapsed time of the experiment was 53 (10.63) minutes. ANOVA testing coupled with Scheffe's pairwise comparison test revealed no significant difference ($\alpha = 0.05$) in mean experimental

Table 1. Sample Characteristics.

Low-Balling	Non-Audit Revenue		
		Sample Size Composition	
Absent	Absent	18	
Absent	Present	17	
Present	Absent	19	
Present	Present	19	
Four of the Big-Five CPA Firms (Participants requested that firm names be disguised)			
Firm-One		19	
Firm-Two		20	
Firm-Three		16	
Firm-Four		18	
Gender			
Male		58	
Female		15	
Total Sample Size (*N*)		73	
		Demographic Variables	
Mean (standard deviation) age		44.60	(6.09)
Mean (standard deviation) years in public accounting		21.07	(5.99)
Mean (standard deviation) years with current firm		18.06	(7.66)
Mean (standard deviation) years as partner		8.70	(5.67)
Percent experience auditing public companies		69.10	(17.36)

times across three of the four treatment conditions; however, mean elapsed time in the condition where both economic rents were present was significantly higher ($\alpha < 0.05$) at 61 (8.29) minutes.

Manipulation Checks

Manipulation check items were used to gauge the effectiveness of the experimental treatments. Participants were asked to respond to two items representing each treatment condition. The manipulation check items for low-balling were:

> I bid this engagement at a loss in order to obtain the client's future audit business (1 = Strongly Disagree, 7 = Strongly Agree).
>
> I used a low-balling strategy to acquire the initial audit engagement (1 = Strongly Disagree, 7 = Strongly Agree).

Correlation between the two low-balling items was 0.69 ($p < 0.01$). The items were summed to obtain a low-balling index. The manipulation check items for potential non-audit revenue were:

In the future, it is likely that my CPA firm will earn non-audit revenues from this client (1 = Strongly Disagree, 7 = Strongly Agree).

In the future, it is likely that non-audit revenues earned from this client will enhance my personal income (1 = Strongly Disagree, 7 = Strongly Agree).

Correlation between the two potential non-audit revenue items was 0.72 ($p < 0.01$). The items were collapsed (summed) into a single non-audit revenue index.

The index means (standard deviations) for low-balling in the absent and present conditions were 4.63 (1.90) and 11.45 (2.05), respectively. A t test indicated a significant mean difference between the low-balling conditions ($t = 14.72$, $p < 0.01$). The index means (standard deviations) for potential non-audit revenue in the absent and present conditions were 4.17 (1.77) and 12.11 (2.07), respectively. A significant mean difference was obtained between the potential non-audit revenue conditions ($t = 17.66$, $p < 0.01$). Accordingly, the manipulations were considered successful.

Means and Correlations

Table 2 (Panel A) presents means and standard deviations of study variables. Table 2 (Panel B) shows the correlation matrix.

Ideally, the research model should be analyzed using structural equations modeling (SEM). However, the relatively small sample size ($n = 73$) precludes the use of SEM testing, as the estimated power of 0.27 is quite low.[5] As an alternative, a path analysis model can be reliably calculated. The statistical package used for the path analysis is called EQS. In the path model, measurement error for the independent variables (i.e. low-balling and potential non-audit revenue) is set to zero, since they are manipulated treatment conditions.

The result of conducting a path analysis on the research model is shown in Fig. 2. The non-significant W statistic of 13.46 ($p = 0.15$) indicates a good model fit (Schumacker & Lomax, 1996; Maruyama, 1997).[6] All standardized path coefficients are significant at $p < 0.01$. When comparing the direct and indirect paths from the independent to dependent variables to the correlation matrix, all residuals fall below 0.07. The relatively low residuals coupled with the non-significant W statistic suggest that the research model adequately fits the data. Further analysis of the path model is offered below.

Hypothesis One

The first hypothesis posits that in the presence, as compared to absence, of low-balling (*H1a*) and potential non-audit revenue (*H1b*), auditors will increase their

Table 2. Process and Consequence Variables.

Panel A: Treatment Means (Standard Deviations)

Low-Balling Potential Non-Audit Revenue	Absent Absent	Absent Present	Present Absent	Present Present
Initial Client Commitment	3.94	4.95	5.11	6.74
	(1.06)	(0.97)	(0.99)	(0.56)
Attribute Ratings	−2.18	0.02	−0.34	2.23
	(1.72)	(1.11)	(1.24)	(0.95)
Importance Weightings	5.61	3.82	3.99	2.38
	(2.07)	(1.43)	(1.39)	(1.02)
Predecisional Distortion Index	−145.76	0.59	−1.32	63.11
	(48.63)	(26.91)	(25.16)	(18.66)
Going Concern Likelihood	41.11	60.59	64.21	85.26
	(16.76)	(22.44)	(24.10)	(13.89)
Revision to Initial Budget Hours	478.89	291.18	276.32	163.16
	(120.39)	(68.65)	(68.64)	(39.02)

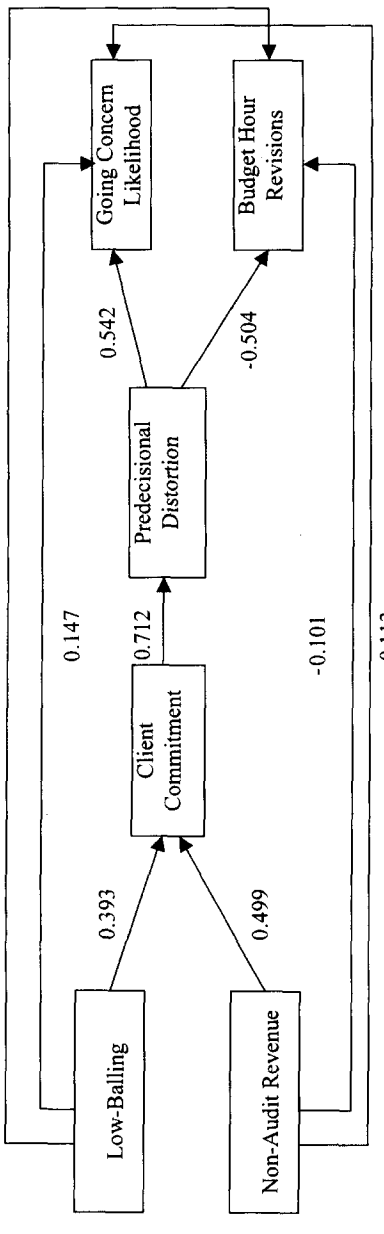
Panel B: Correlation Matrix

Process and Consequence Variables	[1]	[2]	[3]	[4]	[5]	[6]	[7]
[1] Low-Balling*	–						
[2] Non-Audit Revenue*	0.014[&]	–					
[3] Initial Client Commitment	0.601	0.722	–				
[4] Attribute Ratings	0.234	0.302	0.444	–			
[5] Importance Weightings	0.187	0.195	0.257	0.308	–		
[6] Predecisional Distortion Index	0.487	0.562	0.856	0.377	0.265	–	
[7] Going Concern Likelihood	0.360	0.358	0.272	0.351	0.192	0.872	
[8] Budget Hour Revisions	−0.282	−0.343	−0.301	−0.336	0.193	−0.841	−0.407

* Manipulated independent variables.

[&] p-value > 0.10 (p-values for all other correlations < 0.01).

initial client commitment. To test this assertion, an ANOVA model was run using both contingent economic rents as the independent variables and initial client commitment as the dependent variable. As shown in Table 3, initial client commitment was lowest in the absence of contingent economic rents ($m = 3.94$) and highest in the presence of both contingent rents ($m = 6.74$). Initial client commitment in the presence of either potential non-audit revenue ($m = 4.95$) or low-balling ($m = 5.11$) fell between the two extremes. Additional support for H1 is found in the path analysis (see Fig. 2). Specifically, the paths from low-balling (path coefficient = 0.393) and non-audit revenue (path coefficient = 0.499) are positive and significant ($p < 0.01$). Based on the above analyses, *H1a* and *H1b* are supported.



1. The non-significant $W_{(d.f. = 9)}$ statistic of 13.46 ($p = .15$) suggests a good model fit.
2. All paths are significant at $\alpha < .01$.

Model Paths	Direct Path	Indirect Path	Correlation	Residual
Low-Balling to Going Concern Likelihood	.147	.152	.360	.061
Low-Balling to Budget Hour Revisions	-.092	-.141	-.282	-.049
Non-Audit Revenue to Going Concern Likelihood	.113	.193	.358	.052
Non-Audit Revenue to Budget Hour Revisions	-.101	-.179	-.343	-.063

Fig. 2. Path Analysis of Research Model.

Table 3. ANOVA Test Results for Hypothesis One (*H1a* and *H1b*).

Panel A: Initial Client Commitment

Source	d.f.	Sum-Squares	F-Ratio	*p*-value
Low-Balling	1	31.30	37.67	0.001
Non-Audit Revenue	1	39.96	48.09	0.001
Interaction Term	1	1.73	2.08	0.154
Error	69	57.34		
Total (Adj.)	72	131.92		

Results of Scheffe's multiple pairwise comparison test ($\alpha = 0.05$)

Low-Balling	Non-Audit Revenue	Means
Absent	Absent	3.94[a]
Absent	Present	4.95[b]
Present	Absent	5.11[b]
Present	Present	6.74[c]

Different superscripts indicate significant differences at $\alpha = 0.05$. As a result, 3.94 is significantly different from all other means, 4.95 and 5.11 are significantly different from 3.94 and 6.74 but not from each other, and 6.74 is significantly different from all other means.

Hypothesis Two

The second hypothesis (*H2a*) asserts a positive relationship between initial client commitment and predecisional distortion of information. To test this assertion, we again refer to the path analysis (see Fig. 2). As depicted, the path from initial client commitment to predecisional distortion (path coefficient = 0.712) is positive and significant ($p < 0.01$). Accordingly, *H2a* is supported.

Hypothesis *H2b* predicts that the predecisional distortion index, comprised of attribute ratings times importance weightings summed over all informational items, will yield a better model fit than either factor alone. To test this assertion, we ran two additional path analyses wherein we replaced the predecisional distortion index with either attribute ratings or importance weightings. The W fit statistic was highly significant for attribute ratings ($W = 19.43$, $p < 0.02$) and importance weightings ($W = 22.87$, $p < 0.01$), suggesting poor model fits in both instances, thus supporting *H2b*.[7]

Hypothesis Three

The third hypothesis indicates a positive relationship between predecisional distortion and likelihood assessments that the client will continue as a going concern

(*H3a*), and a negative relationship between predecisional distortion and revisions to initial budget hours (*H3b*). Referring to the path analysis (see Fig. 2), the paths from predecisional distortion to going concern likelihood assessments (path coefficient = 0.542) and budget hour revisions (path coefficient = –0.504) are significant ($p < 0.01$). Thus, *H3a* and *H3b* are supported.

The direct effects of low-balling and potential non-audit revenue on the dependent variables are also examined. As indicated earlier, we expected significant direct effects, primarily because contingent economic rents are experimentally manipulated, and going concern assessments and audit plans involve endogenous and exogenous factors not considered in this study. The path model indicates a significant direct path from low-balling to going concern likelihood assessments (path coefficient = 0.147) and budget hour revisions (path coefficient = –0.092). Significant direct paths from non-audit revenue to going concern likelihood assessments (path coefficient = 0.113) and budget hour revisions (path coefficient = –0.101) are also obtained.

Additionally, ANOVA models were used to test the direct impact of contingent economic rents on the dependent variables of interest (see Table 4). As indicated in Table 4, Panel A (Panel B), mean going concern likelihood assessments (budgeted hour revisions) were significantly lowest (highest) in the control condition and highest (lowest) in the presence of both contingent economic rents, with the remaining treatment means falling between the two extremes.

Post-Experiment Client Commitment

Before answering debriefing items, the auditors responded to the following client commitment item:

> Before you examined the client's information, you were asked to respond to the following question, "At this point, I feel very committed to maintaining a long-term relationship with this client." Your response was [*the participant's initial response appeared here*].
>
> Now that you have read more information about your client, please indicate on the scale below the extent to which your initial commitment toward this client has changed (1 = Extremely less committed, 2 = Moderately less committed, 3 = Slightly less committed, 4 = No change, 5 = Slightly more committed, 6 = moderately more committed, 7 = Extremely more committed).

ANOVA testing (see Table 5) reveals a significant difference among treatment conditions, such that higher levels of contingent economic rents appear to trigger greater levels of ex post client commitment. These findings suggest that, in multi-period settings, contingent economic rents may strengthen auditors' predisposition toward maintaining an on-going relationship with the client.

Table 4. ANOVA Test Results for Direct Effect of Independent Variables on Going Concern Likelihood and Revisions to Budgeted Hours.

Panel A: Going Concern Likelihood

Source	d.f.	Sum-Squares	F-Ratio	*p*-value
Low-Balling	1	10,391.39	26.95	0.001
Non-Audit Revenue	1	7,478.80	19.39	0.001
Interaction Term	1	11.30	0.03	0.865
Error	69	26,608.74		
Total (Adj.)	72	44,775.34		

Results of Scheffe's multiple pairwise comparison test ($\alpha = 0.05$)

Low-Balling	Non-Audit Revenue	Means
Absent	Absent	41.11[a]
Absent	Present	60.59[b]
Present	Absent	64.21[b]
Present	Present	85.26[c]

Panel B: Revisions to Budgeted Hours

Source	d.f.	Sum-Squares	F-Ratio	*p*-value
Low-Balling	1	497,585.80	79.11	0.001
Non-Audit Revenue	1	412,138.10	65.52	0.001
Interaction Term	1	25,306.50	4.02	0.049
Error	69	434,006.90		
Total (Adj.)	72	1,377,795.00		

Results of Scheffe's multiple pairwise comparison test ($\alpha = 0.05$)

Low-Balling	Non-Audit Revenue	Means
Absent	Absent	478.89[a]
Absent	Present	291.18[b]
Present	Absent	276.32[b]
Present	Present	163.16[c]

Different superscripts indicate significant differences at $\alpha = 0.05$. As a result, 478.89 is significantly different from all other means, 291.18 and 276.32 are significantly different from 478.89 and 163.16 but not from each other, and 163.16 is significantly different from all other means.

Table 5. ANOVA Test Results for Post-hoc Observations.

Post-Experiment Client Commitment

Source	d.f.	Sum-Squares	F-Ratio	*p*-value
Low-Balling	1	48.06	34.71	0.001
Non-Audit Revenue	1	57.45	41.49	0.001
Interaction Term	1	0.91	0.66	0.420
Error	69	95.53		
Total (Adj.)	72	204.00		

Results of Scheffe's multiple pairwise comparison test ($\alpha = 0.05$)

Low-Balling	Non-Audit Revenue	Means
Absent	Absent	2.39[a]
Absent	Present	3.94[b]
Present	Absent	3.79[b]
Present	Present	5.79[c]

Between-Subject Debriefing

The participants next answered between-subject debriefing items, which were randomized per individual. Wording of the debriefing items and means (standard deviations) are included in Table 6. ANOVA analyses indicated no significant difference across treatment conditions for any of the response items.

The first category of debriefing items assessed the audit partners' beliefs regarding the extent to which they were independent, unbiased, objective, and impartial with respect to the case client. As indicated, mean responses were considerably high for all measured variables. These findings are interesting because if participants honestly believed that they were independent, unbiased, objective, and impartial, then the significantly different going concern likelihood assessments and budget hour revisions across treatment conditions indicate that predecisional distortion of information and biased judgment may operate at the subconscious level.

The second category assessed how the auditors felt about the judgments they had just rendered, which is a reflection of post-decision cognitive dissonance. According to dissonance theory, when conscious deliberate misrepresentation occurs, individuals often experience a tension producing psychological state referred to as cognitive dissonance, which arises when there is an inconsistency between an individual's beliefs and actions (Festinger, 1957, 1964; Brockner, 1992). Accordingly, if the participating auditors were intentionally

Table 6. Means and (Standard Deviations) for Between-Subjects Debriefing Questions.

Perceptions of Independence		
I am independent of this client in all respects.	6.59	(0.93)
My audit judgment with regard to this client is unbiased.	6.56	(1.05)
My audit judgment with regard to this client is objective.	6.41	(1.22)
There are no conflicts of interest impairing my independence on this audit engagement.	6.38	(1.13)
I believe that my audit judgment with respect to this client is impartial.	6.55	(1.01)
Cognitive Dissonance		
I feel comfortable with the judgments I just provided concerning HMW Inc.	6.28	(0.97)
I am uncertain about the judgments I just provided concerning HMW Inc.	2.04	(1.32)
I am uneasy with the judgments I just provided concerning HMW Inc.	2.27	(1.11)
Compensation Salience		
I am satisfied with my compensation for participating in this study, considering the time and effort I just expended.	5.92	(1.38)
Case Instructions		
The case instructions for this study were understandable.	6.03	(1.37)

(1) Scale used for all items: 1 = Strongly Disagree, 2 = Disagree, 3 = Slightly Disagree, 4 = Neither Agree nor Disagree, 5 = Slightly Agree, 6 = Agree, 7 = Strongly Agree,
(2) All means are statistically non-significant ($p > 0.10$) across the between-subjects experimental treatment conditions.
(3) All means are significantly different ($p < 0.01$) from the mid-point of the scale.

misrepresenting the client's financial condition, some degree of cognitive dissonance might be indicated in the contingent economic rents conditions. However, the debriefing items do not indicate varying levels of cognitive dissonance across treatment conditions, as all auditors expressed relatively high levels of comfort with and confidence in their judgments.

The next category, compensation salience, is designed to measure the efficacy of the experimental incentives. Based on the overall mean, the auditors believed that the $50 contribution to their favorite charity was adequate compensation. Finally, the relatively high mean response to the last item suggests that the auditors understood the case instructions.

Within-Subject Debriefing

After answering manipulation checks, between-subject debriefing and demographic items, the auditors responded to a final set of within-subject debriefing

items. The purpose of the within-subject debriefing was to assess the participants' underlying beliefs regarding the extent to which their and other audit partners' judgments would be impaired in light of contingent economic rents. Within-subject assessments of this nature offer participants an opportunity to reflect on the relationship between the primary independent and dependent constructs of interest, unconfounded by the experimental treatment condition to which they were assigned (Kahneman & Tversky, 1996; Libby & Tan, 1999; Tan et al., 2000).

Each participant read four different scenarios, unrelated to the experimental case materials. The four scenarios reflected situations where low-balling (potential non-audit revenue) were absent (absent), absent (present), present (absent) or present (present). After reading each scenario, the participants recorded the extent to which their own judgments might be impaired, as well as how the judgments of other auditor partners in their firm might be compromised. The order of the four scenarios was randomized for each individual.

Item wording and response means (standard deviations) are presented in Table 7. For purposes of statistical testing, the means for "my judgment" items were summed into one index and "other partner judgments" items were summed into another index within each scenario. Within-scenario index means were then tested for significant differences using t tests. Within the first scenario (low-balling – absent; non-audit revenue – absent) the indices were not significantly different from each other ($t = 0.08, p = 0.93$). However, the indices were significantly different within scenarios two (low-balling – present; non-audit revenue – absent) ($t = 16.81$, $p < 0.01$), three (low-balling – absent; non-audit revenue – present) ($t = 17.63$, $p < 0.01$), and four (low-balling – present; non-audit revenue – present) ($t = 37.40$, $p < 0.01$).

Index means across the scenarios were also examined using t tests. The "my judgments" index means were not significantly different across the four scenarios, as all t tests yielded p-values > 0.90. The "other partner judgments" mean index in scenario one was significantly higher than scenarios two ($t = 16.68$, $p < 0.01$), three ($t = 20.37$, $p < 0.01$) and four ($t = 37.83$, $p < 0.01$). In the presence of either contingent economic rent (scenarios two and three), the "other partner judgments" index means were not significantly different from each other ($t = 1.06$, $p = 0.29$). Finally, the "other partner judgments" index mean for scenario four was significantly lower than scenarios two ($t = 12.32$, $p < 0.01$) or three ($t = 15.88$, $p < 0.01$).

Test results suggest that the audit partners do not believe their judgment would be impaired by the presence of contingent economic rents. However, they indicate that the judgments of other partners in their firms might be compromised. In psychological testing of this nature, individuals often record answers that present themselves in the most favorable light (Kerlinger, 1986). However,

Table 7. Means and (Standard Deviations) for Within-Subjects Debriefing Questions.

Scenario One: Low-Balling (Absent), Non-Audit Revenue (Absent)		Mean	S.D.
My audit judgments with respect to this client would be totally:	objective.	6.85a	(0.64)
	unbiased.	6.84a	(0.37)
	impartial.	6.80a	(0.66)
The audit judgments of other partners in my firm would be totally:	objective.	6.93a	(0.25)
	unbiased.	6.78a	(0.73)
	impartial.	6.77a	(0.70)
Scenario Two: Low-Balling (Present), Non-Audit Revenue (Absent)			
My audit judgments with respect to this client would be totally:	objective.	6.81a	(0.59)
	unbiased.	6.81a	(0.49)
	impartial.	6.82a	(0.42)
The audit judgments of other partners in my firm would be totally:	objective.	4.82b	(1.54)
	unbiased.	5.05b	(1.46)
	impartial.	4.77b	(1.65)
Scenario Three: Low-Balling (Absent), Non-Audit Revenue (Present)			
My audit judgments with respect to this client would be totally:	objective.	6.67a	(0.99)
	unbiased.	6.74a	(0.87)
	impartial.	6.78a	(0.69)
The audit judgments of other partners in my firm would be totally:	objective.	5.00b	(1.27)
	unbiased.	4.88b	(1.28)
	impartial.	5.19b	(1.21)
Scenario Four: Low-Balling (Present), Non-Audit Revenue (Present)			
My audit judgments with respect to this client would be totally:	objective.	6.79a	(0.67)
	unbiased.	6.80a	(0.66)
	impartial.	6.84a	(0.65)
The audit judgments of other partners in my firm would be totally:	objective.	3.08b	(1.20)
	unbiased.	3.27b	(1.26)
	impartial.	3.12b	(1.36)

(1) Scale used for all items: 1 = Strongly Disagree, 2 = Disagree, 3 = Slightly Disagree, 4 = Neither Agree nor Disagree, 5 = Slightly Agree, 6 = Agree, 7 = Strongly Agree,

(2) All means are statistically non-significant ($p > 0.10$) across the between-subjects experimental treatment conditions.

(3) Within each scenario, different superscripts indicate significantly different means ($p < 0.01$). That is, means with the same superscript are not significantly different from each other, but means with different superscripts are significantly different from one another.

according to social projection theory, when asked how peers might respond to the same stimuli, respondents often project their own subconscious, or repressed conscious, beliefs onto referent others (e.g. Clement & Krueger, 2000; Mikulincer & Horesh, 1999; Ruvolo & Fabin, 1999; Smith, 1997). Thus, the within-subject debriefing items provide further indication that the presence of contingent economic rents might subconsciously bias the judgment of auditors.[8]

DISCUSSION

The primary purpose of this study is to examine the how and why audit judgment can be biased in the presence of two forms of contingent economic rents: low-balling and potential non-audit revenues. Accordingly, we investigate whether the presence of contingent economic rents establishes an initial predisposition in favor of the audit client, how an initial tendency of this nature can stimulate favorable predecisional distortion of client-related information, and the extent to which predecisional distortion can bias preliminary audit judgment. A fundamental premise of this research is that biased judgment is not always volitional; rather, it often arises from subconscious influences during the collection and evaluation phases of judgment formulation. A secondary objective of this study is to refine the measurement of predecisional distortion of information to include decision weights in an attempt to develop a more robust and predictive distortion metric.

A total of 73 audit partners from four of the Big Five CPA firms participated in a 2 (low-balling: present or absent) by 2 (potential non-audit revenue: absent or present) randomized between-subjects experiment. Research evidence indicates that initial client commitment is highest (lowest) in the presence (absence) of both contingent economic rents, while such commitment falls between the two extremes in the presence of either contingent economic rent. Study findings also reveal a positive relationship between initial client commitment and predecisional distortion of client-related information. Additionally, experimental results suggest that confirmatory processing is positively associated with going concern likelihood assessments and negatively associated with initial budget hour revisions. Finally, statistical testing of the research model demonstrates that the refined predecisional distortion metric developed for this study is more predictive than the measurement technique used in past studies (e.g. Russo et al., 1996, 1998, 2000; Wilks, 2001).

Responses to between-subject debriefing items indicate that auditors in all four treatment conditions believed that their judgments with respect to the case client were objective, unbiased, and impartial. Within-subject debriefing items

were also assessed. After reading each of four scenarios (low-balling and potential non-audit revenue: absent or present) unrelated to the experimental case, auditors again indicated that their client-related judgments would be objective, unbiased, and impartial. Interestingly, in the presence of contingent economic rents, they suggested that the judgments of other audit partners within their firms would be biased. Based on social projection theory, responses to the "other partners" items indicate that the participating auditors are unaware of their own susceptibility to the biasing influence of contingent economic rents, aware of but unwilling to admit to such influence, or some combination thereof.

The current study is limited by external validity threats common to laboratory experiments of this nature. However, using practising audit partners as participants and a reality-based case provides some degree of generalizability to the results. The study is also limited in that participants are not asked to render a final audit decision. Results of pilot testing, with a different group of audit partners, suggested that there is too little information in the case for auditors to provide opinions, particularly with respect to going concern; hence, likelihood assessments would make more sense to the respondents. Thus, the extent to which professional responsibility, legal concerns, accountability, and other factors might affect the auditors' final opinions is unknown. For instance, Smith and Kida (1991) suggest that auditors' judgmental biases of this nature do not always result in a pro-client opinion due to their application of conservatism and professional skepticism. Another limitation of this study concerns the nature of the contingent economic rents manipulations. That is, the presence of low-balling and potential non-audit revenue might lead auditors to believe that the client must be financially viable, else, why would the firm accept the client in the first place. We believe that differing client risk perceptions of this nature are a natural confound in the business world, and that it is not possible to experimentally disentangle such a confound in an experimental setting. In fact, this very confound has ignited heated debates regarding auditor independence related to contingent economic rents. Hence, the extent to which this issue might have affected study results is unknown. Finally, with respect to the debriefing items, participant responses may not be reflective of their true beliefs regarding independence related issues; rather, the auditors may have responded in a way that placed them in the most favorable and accepted light. Thus, whether confirmatory processing is truly subconscious cannot be determined with absolute certainty in this experiment.

As stated by Bazerman et al. (1997, p. 91), "The assumption of deliberativeness is important because it implies that any tendency toward bias can be potentially rectified by moral suasion and/or the threat of sanctions." However, if confirmatory processing is wholly or partially subconscious, such remedies

may be ineffective. Therefore, future research in the area of auditor independence should focus on developing self-awareness techniques that audit partners and firms can use to identify situations where independence may be impaired. Self-awareness procedures of this nature could be integrated into firm training sessions and auditor decision aids. Also, it would be fruitful to investigate the impact of other possible threats to independence, such as the size and nature of equity stakes with clients and the length of the auditor-client relationship. In the final analysis, the accounting profession should take proactive steps to identify and deal with threats to auditor independence.

NOTES

1. We recognize that the going concern likelihood assessments obtained in the current experiment reflect a preliminary phase of the audit, such as 'understanding the client'. We further acknowledge that an auditor's final opinion in this regard might not be affected by such initial beliefs due to a host of individual and contextual factors, such as professional responsibility, legal concerns, accountability and conservatism.

2. The initial budget hours selected for the case were derived from actual audit files from two Big-five CPA firms that had audit clients in the same industry and of the approximate size as the case client in this study. After reviewing six files of companies that had no serious audit issues, we set the initial budget hours 10% below the low end of the observed spectrum, so that all auditors would want to revise them upward, to some extent, particularly since the case client was in financial trouble.

3. During pilot testing, the researchers first had the auditors evaluate each piece of information separately, in serial order. However, pilot participants did not like this idea, since many of the client-related informational items are inter-dependent. Rather, the pilot participants requested that they be able to evaluate all pieces together, after a thorough review of available information. In the Russo et al. (1996, 1998, 2000) studies, the object attributes were independent, hence serial evaluation was appropriate. Our procedure in this regard is consistent with the interdependent attribute evaluation method suggested by Carlsson and Fuller (1995).

4. For instance, if a participant rated a piece of information as +2 on the rating scale and 5 on the weighting scale, the predecisional distortion metric for that information item would be +10. The distortion metric for each of the 12 pieces of information is summed to obtain the overall index value for a given participant. The predecisional distortion metric for a single informational item can range from –21 (-3×7) to +21 ($+3 \times 7$) and the possible index range for a given participant over all 12 pieces of information is from –252 (-21×12) to +252 ($+21 \times 12$).

5. SEM power was tested via the procedure described in MacCallum et al. (1996).

6. The W statistic is obtained by comparing the observed and reproduced correlation matrices. The W statistic accounts for both the number of variables in the model and sample size, and it approximates the X^2 distribution. A non-significant W statistic suggests a good model fit.

7. Workshop participants at Cornell University mentioned that having participants record attribute ratings and importance weightings side-by side (as in the current

experiment) might serve to allocate attribute ratings (as used in Russo et al. [1996, 1998, 2000]), between the ratings and weightings. As a result, we conducted a second experiment using the same case materials as in the main experiment of this paper (where both low-balling and non-audit revenues were absent). A total of 63 audit seniors and managers representing three of the big-five CPA firms participated in the experiment wherein participants assessed attribute ratings only, importance weightings only, or both simultaneously (as in the current experiment). Regression results (adjusted model $R^2 = 0.70$) indicate that the predecisional distortion index (ratings times weightings summed over all informational items) was significant ($t = 6.33$, $p < 0.01$), and attribute ratings ($t = 1.59$, $p = 0.13$) and importance weightings ($t = 2.01$, $p = 0.22$) were both non-significant. Hence, the predecisional distortion metric was significantly more predictive than either factor alone, implying that assessing both ratings and weightings simultaneously does not merely allocate attribute ratings between the two factors.

8. Participants' responses to the within-subject items might also reflect their repressed conscious beliefs with respect to the influence of contingent economic rents on independence. Meaning, they might be aware of their own susceptibility to biased judgment in this regard, but are unwilling to admit such beliefs due to social unacceptability. Hence, we recognize that both subconscious and conscious beliefs may become manifest through social projection of this nature.

REFERENCES

Antel, R. (1984, Spring). Auditor independence. *Journal of Accounting Research*, 1–20.

Baber, W., Brooks, E., & Ricks, W. (1987, Autumn). An empirical investigation of the market for audit services in the public sector. *Journal of Accounting Research*, 293–305.

Bamber, M. E., Ramson, R. J., & Tubbs, R. M. (1997). An examination of the descriptive validity of the belief-adjustment model and alternative attitudes to evidence in auditing. *Accounting, Organizations & Society*, *122*(3, 4), 249–268.

Bazerman, M., Morgan, K., & Loewenstein, G. (1997, Summer). The impossibility of auditor independence. *Sloan Management Review*, 89–94.

Bedard, J. C., & Biggs, S. F. (1991). Pattern recognition, hypotheses generation, and auditor performance in an analytical task. *The Accounting Review*, *66*(3), 622–642.

Brockner, J. (1992). The escalation of commitment to a failing course of action: Toward theoretical progress. *Academy of Management Review*, *17*, 39–61.

Brown, C., Peecher, M., & Solomon, I. (1997). Auditors' hypothesis testing in diagnostic inference tasks. *Journal of Accounting Research*, *37*(1), 1–26.

Calegari, M., Schatzberg, J., & Sevick, G. (1998). Experimental evidence of differential auditor pricing and reporting strategies. *The Accounting Review*, *73*(2), 255–276.

Carlsson, C., & Fuller, R. (1995, March). Multiple criteria decision making: The case for interdependence. *Computers and Operations Research*, *22*(3), 251.

Chapman, G. B., & Elstein, A. S. (2000). Cognitive processes and biases in medical decision making. In: G. B. Chapman & F. A. Sonnenberg (Eds), *Decision Making in Health Care: Theory, Psychology, and Applications* (pp. 3–438). New York, NY: Cambridge University Press.

Church, B. K. (1990). Auditors' use of confirmatory processes. *Journal of Accounting Literature*, *9*, 81–112.

Church, B. K. (1991). An examination of the effect that commitment to a hypothesis has on auditors' evaluations of confirming and disconfirming evidence. *Contemporary Accounting Research, 7*(2), 513–534.

Church, B. K., & Schneider, A. (1993). Auditor's generation of diagnostic hypotheses in response to a superior's suggestion: Interference effects. *Contemporary Accounting Research, 10*(1), 330–350.

Clement, R. W., & Krueger, J. (2000). The primacy of self-referent information in perceptions of social consensus. *British Journal of Social Psychology, 39*(2), 279–299.

Cloyd, C. B., & Spilker, B. C. (1999). The influence of client preferences on tax professionals' search for judicial precedents, subsequent judgments and recommendations. *The Accounting Review, 74*(3), 299–322.

DeAngelo, L. (1981, August). Auditor independence, 'low balling,' and disclosure regulation. *Journal of Accounting and Economics*, 113–128.

Elliott, A. J., & Devine, P. G. (1994). On the motivational nature of cognitive dissonance. *Journal of Personality and Social Psychology, 67*, 382–394.

Farmer, T., Rittenberg, L., & Trompeter, G. (1987, Fall). An investigation of the impact of economics and organizational factors on auditor independence. *Auditing: A Journal of Practice and Theory*, 1–14.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Festinger, L. (1964). *Conflict, Decision, and Dissonance*. Stanford, CA: Stanford University Press.

Fischhoff, B. R., & Beyth-Marom, R. (1983, July). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, 239–260.

Fishbein, M. (1983). An investigation of the relationships between beliefs about an object and the attitude toward that object. *Human Relations, 16*, 233–240.

Francis, J., & Simon, D. (1987, January). A test of audit pricing in the small-client segment of the U.S. audit market. *The Accounting Review*, 145–157.

Frey, D. (1986). Recent research on selective exposure to information. In: L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (pp. 41–80). New York, NY: Academic Press.

Gorman, M. (1986, February). How the possibility of error effects falsification on a task that models scientific problem solving. *British Journal of Psychology*, 85–96.

Jarvis, W., Petty, B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of Personality and Social Psychology, 70*(1), 172–194.

Johnson, L. (1993). An empirical investigation of the effects of advocacy on preparers' evaluations of judicial evidence. *Journal of the American Taxation Association, 15*(1), 1–22.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*(3), 582–601.

Keeney, R. L., & Raiffa, H. (1976). *Decisions with Multiple Objectives, Preferences and Value Tradeoffs*. New York, NY: John Wiley.

Kerlinger, F. N. (1986). *Foundations of Behavioral Research* (3rd ed.). Philadelphia, PA: Holt, Rinehart and Winston, Inc.

Kida, T. (1984, Spring). The impact of hypothesis-testing strategies on auditors' use of judgment data. *Journal of Accounting Research*, 332–340.

Knapp, M. C. (1985). Audit conflict: An empirical study of the perceived ability of auditors to resist management pressure. *The Accounting Review, 60*(2), 202–211.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498.

Lee, C., & Gu, Z. (1998, October). Low balling, legal liability, and auditor independence. *The Accounting Review, 73*, 533–553.

Libby, R. (1985). *Accounting and human information processing: Theory and applications.* Englewood Cliffs: NJ: Prentice-Hall, Inc.

Libby, R., & Tan, H. T. (1999). Analysts' reaction to warnings of negative earnings surprises. *Journal of Accounting Research, 37*(2), 415–435.

Lord, C., Ross, L., & Lepper, M. (1979, November). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 2098–2109.

Lweicka, M. (1998). Conformation bias: Cognitive error or adaptive strategy of action control? In: M. Kofta & G. Weary (Eds), *Personal Control in Action: Cognitive and Motivational Mechanisms* (pp. 233–258). New York, NY: Plenum Press.

MacCallum, R., Browne, M., & Sugawara, H. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130–149.

Magee, R., & Tseng, M. (1990, April). Audit pricing and independence. *The Accounting Review*, 315–336.

Maruyama, G. (1997). *Basics of Structural Equation Modeling.* Thousand Oaks, CA: Sage Publications.

McKinley, S., Pany, K., & Reckers, P. (1985, Autumn). An examination of the influence of CPA firm type, size and MAS provision on loan officer decisions and perceptions. *Journal of Accounting Research*, 887–896.

McMillan, J. J., & White, R. A. (1993). Auditors' belief revisions and evidence search: The effect of hypothesis frame, confirmation bias, and professional skepticism. *The Accounting Review, 68*(3), 443–465.

Messick, D. M., & Sentis, K. P. (1979). Fairness and preference. *Journal of Experimental Social Psychology, 15*, 418–434.

Mikulincer, M., & Horesh, N. (1999). Adult attachment style and the perception of others: The role of projective mechanisms. *Journal of Personality & Social Psychology, 76*(6), 1022–1034.

Montgomery, H. (1983). Decision rules and the search for dominance structure: Towards a process model of decision making. In: P. Humphreys, O. Seveson & A. Vari (Eds), *Analyzing and Aiding Decision Processes* (pp. 324–348). Amsterdam, The Netherlands: North-Holland.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(22), 175–220.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings in human judgment.* Englewood Cliffs, NJ: Prentice Hall.

Pany, K., & Reckers, P. (1984, Spring). Non-audit services and auditor independence – A continuing problem. *Auditing: A Journal of Practice and Theory, 3*, 89–97.

Pany, K., & Reckers, P. (1988). Auditor performance of MAS: A study of its effects on decisions and perceptions. *Accounting Horizons, 2*(2), 31–38.

Pyszczynski, T., & Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. In: L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (pp. 297–340). New York, NY: Academic Press.

Russo, J. E., Medvec, V., & Meloy, M. (1996). The distortion of information during decisions. *Organizational Behavior and Human Decision Processes, 66*(1), 102–110.

Russo, J. E., Meloy, M., & Medvec, V. (1998, November). Predecisional distortion of product information. *Journal of Marketing Research, 34*, 438–452.

Russo, J. E., Meloy, M. G., & Wilks, T. J. (2000). Predecisional distortion of information by auditors and salespersons. *Management Science, 46*(1), 13–27.

Ruvolo, A. P., & Fabin, L. A. (1999). Two of a kind: Perceptions of own and partner's attachment characteristics. *Personal Relationships, 6*(1), 57–79.

Schatzberg, J. (1990, October). A laboratory market investigation of low balling in audit pricing. *The Accounting Review*, 337–362.

Schatzberg, J., & Sevcik, G. (1994, Summer). A multiperiod model and experimental evidence of independence and "low balling". *Contemporary Accounting Research*, 137–174.

Schatzberg, J., Sevcik, G., & Shapiro, B. (1996, Supplement). Exploratory experimental evidence on independence impairment conditions: Aggregate and individual results. *Behavioral Research in Accounting, 8*, 173–195.

Schumacker R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Schockley, R. (1981, October). Perceptions of auditors' independence: An empirical analysis. *The Accounting Review*, 785–800.

SEC (Securities and Exchange Commission) (2000). *Final rule: Revision of the commission's auditor independence requirements* (http://www.sec.gov/rules/final/33–7919.htm).

Sevenson, O. (1992). Differentiation and consolidation theory of human decision making: A frame of reference for the study of pre- and post-decision processes. *Acta Psychology, 80*, 143–168.

Shafir, E. (1993). Choosing vs. rejecting: Why some options are both better and worse than others. *Memory & Cognition, 21*, 546–556.

Simon, D., & Francis, J. (1988, April). The effects of auditor change on audit fees: Tests of price cutting and price recovery. *The Accounting Review*, 255–269.

Simunic, D. (1984). Auditing, consulting, and auditor independence. *Journal of Accounting Research, 22*, 679–702.

Smith, E. (1997). Private selves and shared meanings: Or forgive us for our projections as we forgive those who project into us. *Psychodynamic Counseling, 3*(2), 117–131.

Smith, J. F., & Kida, T. (1991). Heuristics and biases: Expertise and task realism in auditing. *Psychological Bulletin, 109*(3), 472–489.

Tan, H. T., Libby, R., & Hunton, J. E. (2000). Analysts' ReacPreannouncement Strategies. *Journal of Accounting Research*, accepted and forthcoming.

Titard, P. (1971, July). Independence and MAS opinions and financial statements users. *Journal of Accountancy*, 47–52.

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Wilks, J. T. (2001). Predecisional distortion of evidence as a consequence of real-time audit review. Working Paper, Brigham Young University.

Zeleny, M. (1984). *MCDM past decade and future trends – A source book of multiple criteria decision making*. Greenwich, CT: JAI Press.

AUDITORS' MEMORY FOR DOCUMENTED EVIDENCE

R. David Plumlee, Brad Tuttle and Cindy Moeckel

ABSTRACT

This study examines the relationships among auditors' evidence documentation mode, their memories for the documented evidence, and judgment. We find that prepared checklists are completed faster and require less time reviewing evidence, yet the risk assessments did not differ from those using written documentation. Those who used written documentation tended to include a higher proportion of positive fraud indicators. Memory measures, shows that documented items are correctly recognized significantly more often than undocumented items. We also find that the recognition rate for those using written documentation is significantly lower than for those using a checklist. Ex post analyses find that checklist users have better recognition of positive items and, as their memory for positive indicator increases, so does their perceived likelihood of fraud.

INTRODUCTION

In 1998, Chairman Arthur Levitt of the SEC (Levitt, 1998) called for improved external auditing. He cautioned, "We cannot permit thorough audits to be sacrificed for re-engineered approaches that are efficient, but less effective".[1] A significant underlying aspect of any effort to improve external auditing is the need to understand how audit judgments are formed and how quality is affected

Advances in Accounting Behavioral Research, Volume 5, pages 51–75.

ISBN: 0-7623-0953-9

by choices of audit technology and technique. Among possible choices about audit technology and procedures are the use of decision aids, staff rotation, and evidence documentation mode (e.g. checklists or memoranda). This study contributes to the effort called for by Chairman Levitt by investigating the possible judgment effects of documentation mode. We explore the links among the means by which auditors document fraud-related evidence, remember documented evidence, and render judgments – all within the context of choice among audit technology and procedures available to auditors. If the means by which auditors document evidence affects their judgments, then the documenting mode could be an overlooked factor with unknown impacts on the audit process.

This study explores several possible effects of documentation mode as a context for consideration of the larger issue of auditors' choice of techniques. Assessment of the likelihood of management fraud is a highly judgmental area, with significant ramifications not only for each engagement, but also for the auditing profession as a whole, clients or potential clients, and the economy. Currently, SAS No. 82 (AICPA, 1997) explicitly requires auditors to document their responses to the fraud risk factors present in an engagement, but it is silent as to the means of documentation auditors should use. Lacking definitive guidance regarding the possible effects of this choice, current practice varies widely and could be assisted by examinations of alternative approaches (Shelton et al., 1999, pp. 15, 19). Perhaps most important, because lists of specific fraud indicators have been refined (Ramos & Lyons, 1997) and in use at least since SAS No. 53 was promulgated, it is possible to compare memoranda to existing well-developed checklist items. Thus, results showing the effects of implementation of the checklist documentation mode can demonstrate pitfalls or advantages attached to its use, compared to the memorandum mode, and help determine whether efforts to refine checklists to this level in other areas are prudent.

Previous audit research that focuses directly on documentation finds that documentation mode affects the amount and type of information documented (Purvis, 1989). It also finds that, while documentation using questionnaires improves comprehensiveness and uniformity of documentation, not using a questionnaire may result in a detectable effect on auditors' fraud risk assessments in some circumstances (Pincus, 1989). However, unlike our study, the previous research did not include the possibility that evidence documentation influences the memory structures that are used when judgments are formulated.

We focus our research on the individual auditor who initially evaluates evidence and makes a preliminary assessment of fraud risk. At the same time, we acknowledge that many critical audit decisions are not made by individuals and the review process is an inherent safeguard in the audit process against the biases of individual auditors (e.g. Libby & Trotman, 1993). A finding that moti-

vates our investigation is that, in certain circumstances, auditors form memories of evidence that may not be accurate (Moeckel, 1990). Nevertheless, they often choose to rely on their memories despite the availability of documentation (Moeckel & Plumlee, 1989). Also, what auditors remember about an audit has been shown to affect their subsequent audit judgments (e.g. Choo & Trotman, 1991; Nelson et al., 1995). Because auditors often rely on their memories, which affects their judgments, differences in memories for evidence induced by different modes of documentation may lead to differences in subsequent judgments. Thus, examining the impact of documentation on preparers' memories and judgments is an important step in understanding the role of documentation in fraud risk assessments, or any other audit judgment.

With the participation of 67 practicing auditors, we conducted an experiment to address questions regarding the effects of documentation mode on auditors' memories and judgments. The auditors documented evidence related to the risk of management fraud for a hypothetical client. The auditors were directed to document specific items of evidence using either a checklist or written notes. Following a delay, the auditors completed a recognition test for the evidence that they had reviewed and documented, after which they assessed the risk of management fraud for the client. The results show that those using checklists spend significantly less time documenting and evaluating evidence. Documented items are recognized correctly more often than undocumented items, and the recognition rate for those using written documentation is significantly lower than for those using a checklist. Ex post analyses comparing the recognition rates between positive and negative fraud indicators for the two documentation conditions find that checklist users have better recognition of items indicating the presence of fraud, and those who use written documentation have a lower recognition rate for indicators that do not suggest fraud.

The following section describes relevant previous research and develops the expectations tested. The third section explains the experimental design and methods used to collect data. The analysis and results are reported in the fourth section, and a discussion of the results and their implications is included in the last section of the paper.

BACKGROUND LITERATURE AND HYPOTHESIS DEVELOPMENT

Characterization of Two Alternative Forms of Documentation

Checklists and memoranda are two basic types of documentation auditors could select. The structure inherent in documentation with a checklist affects many

important processes. By including some items and excluding others, checklists predetermine the items to be considered in a given task. They explicitly set the span of items to be considered. They implicitly aid in evaluation of items and assure that all the bases have been covered. Thus, use of checklists enables the user to defer making judgments about individual items or sets of items. Finally, the act of documentation itself is as simple as making a tick mark. Thus, checklists carry the potential of being completed in a mechanistic fashion, with little attention to the overall meaning of larger sets of items.

On the other hand, when using a memorandum, the documentation task provides general direction, but no particular structure. In order to do the work to support documentation with a memorandum, the auditor must retrieve a suitable task template from memory, or construct one to fit the situation as evidence gathering and evaluation proceed. The template guides those documenting with a memorandum to figure out where to look and how deeply in each area, then to evaluate the relative importance of individual items of evidence encountered and how well each fits the template. These evaluations require elaboration and attention to extract meaning from the data. At the very least, preliminary judgments must be made concurrently on whether or not to document an item and, if documented, what emphasis and interpretation to place on the item. Finally, the auditor must form sentences and ensure s/he has communicated what was intended.

Previous research supplies some empirical evidence about the nature of these different approaches. Purvis (1989) instructed auditors to document internal accounting control systems using one of three formats: checklist, flowchart, or narrative memorandum. Participants submitted information requests to the experimenter and received answers on cards. Purvis found that checklist participants requested significantly more evidence than memorandum participants did (presumably as required by the checklist), while they failed to obtain other particular information that memorandum participants repeatedly requested. The failure of checklist users to consider important evidence not on the checklist is consistent with concerns about mechanistic processing in which the structure of the checklist, rather than the auditor's understanding, drives information search.

Pincus (1989) used two conditions in a task that required participants to "assess the chance that material fraud exists." Participants in one condition completed a 73-item red flag questionnaire, while participants in the other condition made "lists of facts or impressions" (p. 158) that they considered relevant in assessing the possibility of material fraud at the client. The list/impression condition corresponds to a memorandum documentation mode in all aspects, except for formally communicating conclusions using sentences. Pincus found

that auditors in the list/impressions condition were much more selective in their documentation than Checklist participants. When Pincus examined the total number of items included by each group, she found that the list/impression participants documented only about one-sixth as many items as were on the questionnaire. In addition, list/impression participants documented over seven times as many positive fraud indicators as no-fraud indicators, compared to a one-to-one ratio established automatically for checklist participants. Thus, list/impression participants focused more narrowly and deeply than did checklist participants.

Purvis (1989) and Pincus (1989) examined the influence of documentation modes on the nature of the evidence that auditors document. However, when Pincus sought to find a link between what was documented and subsequent judgments, her results were mixed. As part of her study, Pincus used an additional between-subjects factor, i.e. two different cases that were based on an actual firm. One case contained a material fraud while the other case included the same set of facts, except that the effects of the fraud were remedied. For the case where no fraud existed, Pincus found no effect for documentation mode on the fraud risk assessments. In the fraud case, she found that list/impression participants were moderately higher in their risk assessments, which was contrary to the hypothesized relationship.[2] Thus, while Pincus found differences between the amount and nature of the items documented, such differences did not translate into the expected differences in fraud risk assessments.

Other accounting and auditing research shows that judgments are related to recollections from memory (e.g. Choo & Trotman, 1991; Moser, 1989; Nelson et al., 1995). Because Pincus did not measure auditors' memories of the documented evidence, her results do not offer any insight into the possibility that the documentation process affects auditors' memories for evidence. Focusing on recognition following an evidence-related judgment, Ricchiute (1997) shows that subordinate auditors exhibit a confirmation bias toward recognizing facts that are consistent with their judgment. However, the participants in Ricchiute's study did not document the evidence that they were asked to recognize, nor were different modes of documentation used. Our study contributes to understanding the documentation process by including measures of what auditors remember from the evidence they document, as well as measures of the items actually documented and judgments based on the evidence.

Effects of Mode on Memory Encoding and Retrieval

We assume that memory components include short-term memory (STM) and long-term memory (LTM), and two basic cognitive processes – encoding

and retrieval. The act of documenting an informational item causes it to be considered more carefully; thus, documented evidence receives more elaboration in STM, which in turn makes it more retrievable later (Craik & Lockhart, 1972). Comparing auditors' recognition rates for documented items with their rates for undocumented items can assess the effects of the act of documentation on auditors' memory for audit evidence. The predicted effect is as follows.

H1: Regardless of documentation mode, recognition rates will be higher for documented evidence than for undocumented evidence.

We expect the cognitive processing differences induced by documentation mode will in turn condition what an auditor will be able to retrieve from memory. For auditors writing notes, the evidence that they should consider is not explicitly set out, as it is when a checklist is used. Hence, the note writing auditors must generate from LTM a specific template for the types of items that will be documented, and then determine whether each item should be included in the notes. Auditors using notes must, therefore, process the meaning of each item in the given task context, then process further to formulate exactly how to record their observations about the subset of items selected for inclusion. Thus, items documented using notes will receive more cognitive processing than items documented using checklists. Since auditors using checklists engage in neither the evaluative process of deciding which items to document nor in the additional elaboration required to execute the documentation step, documentation with checklists does not involve the degree of elaboration that documentation using notes does. Thus, the following hypothesis is formulated.

H2: Those who document with notes will have a higher recognition rate for documented items than those who use a checklist.

The cognitive effort necessary to document evidence using notes is assumed to require substantial attention and STM capacity for those items documented; hence, little attention and STM capacity is used for undocumented items. Conversely, auditors who use the checklist mode devote less attention and STM capacity to documented items; therefore, they can focus relatively more attention and STM to undocumented items. This leads to the following hypothesis.

H3: Those who use a checklist will have a higher recognition rate for undocumented items than those who use notes.

Documentation Mode and Judgment

Previous studies that used information search techniques to examine auditors' search for fraud-related information also examined fraud risk assessments. As

described above, Pincus (1989) compared the use of questionnaires and open-ended lists of facts or impressions. She found that, in a case representing a no-fraud client, there was no difference in the fraud assessments between those who used the questionnaire and those who listed impressions. However, in a case with a high fraud risk, those who listed their impressions rated the likelihood of fraud significantly higher than did those who used the questionnaire. As noted above, list/impression participants focused more on positive indicators of fraud during documentation.

Zimbelman (1997) compared auditors using holistic or decomposed risk assessment scales. When he examined their risk assessments, he found mixed results. For assessments of intentional misstatements, both groups differed between high and low risk cases. For assessments of unintentional misstatements, the decomposed condition did not differ across risk levels.

Thus, previous research is equivocal about the nature of the possible relationship between documentation mode and judgment. Our examination focuses directly on the nature of this possible relationship. Based on our earlier characterization of the evidence accumulation, processing, and retrieval effects of documentation mode, we formulate two expectations. First, we expect that judgments by those using a checklist will be related to whatever were the implications of the actual items that happened to be included on the checklist; while we expect that judgments by those preparing notes would tend to be more idiosyncratic, reflecting individual choice of emphasis, which results in greater judgment variance. Second, we speculate that, in general, those using notes will tend toward conservatism and render judgments aimed at minimizing the possibility of errors with the worst decision consequences. In this particular fraud assessment task, we expect that those who document using notes will tend to judge higher likelihood of fraud because of the nature of the information search induced by their documentation mode (Pincus 1989). We offer the following hypothesis about the effect of mode of documentation on judgment.

H4: Judgments by those documenting using notes will exhibit higher fraud likelihood assessments and greater variances than judgments by those documenting using checklists.

EXPERIMENTAL METHODS

Experimental Materials and Procedures

A regional human resource (HR) director for an international accounting firm and five of her fellow regional HR directors agreed to participate providing

study participants. Each of the HR directors was sent 24 packets containing a computer disk with the experimental materials and instructions on how to run the program. The regional HR directors were instructed to select staff auditors in their region with two or three years of audit experience to participate using their personal computers. The disk each participant received was labeled with the firm's name and had his or her region's HR director listed on the label. The participants were instructed not to talk about the study until after the disks were returned to their HR director. After receiving the disks from the participants, the HR directors returned them to the researchers. One regional HR office did not distribute the disks due to an oversight. Of the remaining 96 disks, 69 were returned yielding a response rate of 72%. Two of the 69 disks returned were incomplete, resulting in a useable response rate of 70%.

The experimental materials were contained entirely on the computer disks. The materials consisted of a HyperCard program that presented all materials as a series of screens. The sequence of tasks contained in the program is shown in Fig. 1. As shown in Fig. 1, participants in the two documentation conditions generally viewed the same screens except for a filler task, which was performed by checklist condition participants.

The experimental task consisted of two phases: one in which participants reviewed and documented evidence regarding a hypothetical client and another where they performed a recognition task and made judgments regarding the likelihood of fraud. In Phase I, the introduction screen explained each participant's role as the "person responsible for the initial assessment of the risk of management fraud" and described how to document the evidence evaluated, depending on the participant's experimental condition.[3] Next came the screens on which participants would document the evidence, followed by a short narrative (approximately 150 lines long presented in a text box) that described the presence or absence of positive indicators of fraud for the hypothetical audit client. Of 24 items used later on the recognition test, six were stated positively as indicators of fraud, and 18 were stated negatively as raising no "red flag" for the presence of fraud.

After reviewing the narrative, the participants returned to the documentation screens and documented the evidence contained in the narrative. The documentation was constructed to cover three types of information that indicate either a positive or negative tendency toward management fraud: client conditions, management attitudes, and management motivations. All participants received instructions to document the same twelve conditions, attitudes, or motivations (Loebbecke & Willingham, 1988; Eining, Jones & Loebbecke, 1997). Participants were told that they could take as much time as needed and were allowed to move freely between the evidence and the documentation screens.

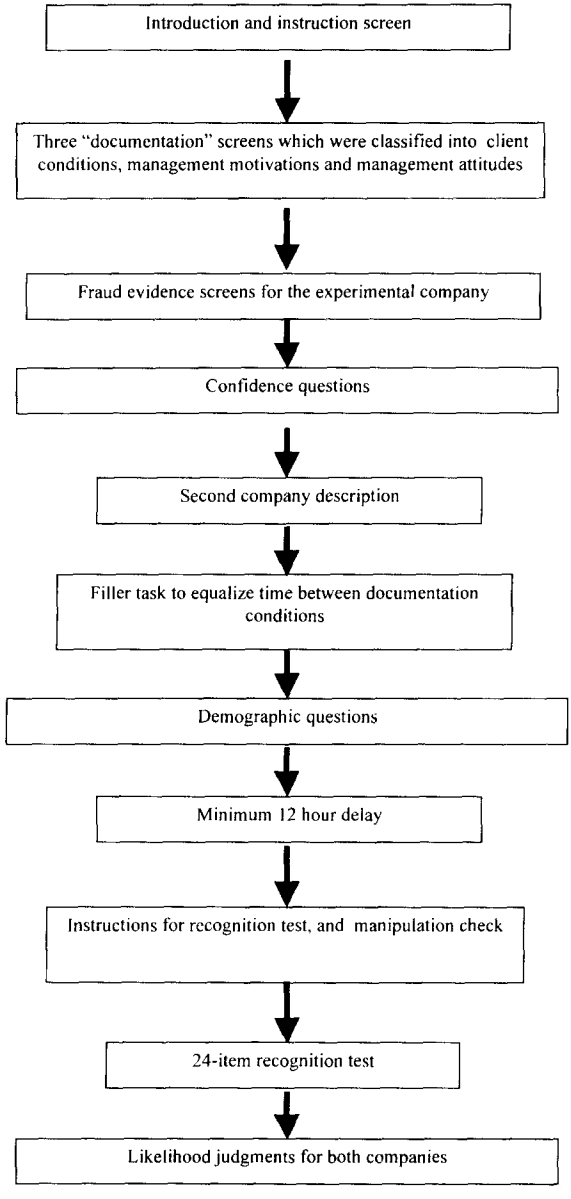


Fig. 1. Sequence Of Experimental Tasks For Both Notes And Checklist Conditions.

Documentation mode (MODE) differed across two sets of participants who were randomly assigned to the two groups. In the checklist condition, participants documented evidence using a 12-item checklist of questions answered by clicking a "yes" or "no" button on the screen. The questions covered fraud indicators that participants were told came from SAS No. 53[4] and research conducted by an international accounting firm. The 12 items included on the checklist are shown in Exhibit 1 (along with 12 additional items on the recognition test that participants were not asked to document).

Participants in the notes condition were instructed to write notes that they might use later to construct a memorandum to document evidence regarding the likelihood of management fraud. These notes were typed into the computer by the participants and stored automatically by the HyperCard program. In order that the notes task paralleled the checklist task, instructions on each documentation screen in the notes condition contained sets of phrases that prompted notes participants to address the same twelve issues covered by the checklist. For example, one of the documentation screens for the notes participants included instructions with the following statement: "Evidence of motivations that might lead to management fraud includes: client experiencing rapid growth, profitability relative to the industry inadequate or inconsistent." There were two corresponding questions on the checklist: "Is the client in a period of rapid growth?" and "Is the client's profitability relative to its industry inadequate or inconsistent?" Each of the 12 questions on the checklist had a corresponding prompt in the Notes conditions' instructions, and every prompt in the instructions given to those who wrote notes was derived from a corresponding question on the checklist. Also, with the following questions, notes participants were specifically cued to consider both positive and negative indicators of fraud: "In the box below, jot down your notes about the motivations [conditions; attitudes] that are present to support a high likelihood of fraud on this client. What motivations [conditions; attitudes] are absent to support a low likelihood of fraud?" Thus, the only substantive difference in the two MODE conditions was the means of documentation.

Once the participants were satisfied with their documentation, they went to the next computer screen and rated their confidence in their future memory-based performance using an 11-point scale. They were asked, "If you were provided some statements about detailed facts or situations found in the evidence you just read, what percentage of the time would you be able to correctly state whether the statements were supported by the evidence?" Participants chose one of eleven buttons, labeled "0%" through "100%" in 10% increments.[5]

Next, the participants reviewed (but did not document) evidence regarding a second client. The second client case was used to reduce a ceiling effect on

Exhibit 1. Recognition Test Items, and whether: (1) they were on the Checklist; and (2) the Evidence Supports the Possibility of Fraud.

Questions used as the bases for the recognition test items	Included on checklist ?	Evidence affirms?
Are management operating and financial decisions dominated by a single person or a few persons acting in concert?	yes	yes
Does management place undue emphasis on meeting earnings projections or other quantitative targets?	yes	yes
Have managers recently entered into collusion with outsiders?	no	no
Does your experience with management indicate a degree of dishonesty?	yes	no
Does management display a propensity to take undue risks?	no	no
Does management display significant disrespect for regulatory bodies?	yes	no
Have managers lied to the auditors or been overly evasive in responses to audit inquiries?	yes	no
Do client personnel display significant resentment of authority?	no	no
Is managements' attitude toward financial reporting unduly aggressive?	no	no
Do key managers exhibit strong personality anomalies?	yes	no
Is this a new client?	yes	no
Is there a need to cover up an illegal act?	no	no
Does the client have a weak control environment?	yes	no
Are there frequent and significant difficult-to -audit transactions of balances?	no	yes
Is the client a public company?	no	no
Is a significant amount of judgment involved in determining the total of an account balance or class of transactions?	no	yes
Have there been instances of material management fraud in prior years?	no	no
Is the client's organization decentralized without adequate monitoring?	yes	no
Is the client in a period of rapid growth?	yes	no
Does the client have solvency problems?	yes	no
Does a conflict of interest exist involving the client entity and/or its personnel?	no	no
Do accounting personnel exhibit inexperience or laxity in performing their duties?	no	no
Is the client confronted with adverse legal circumstances?	yes	yes
Is the client's profitability relative to its industry inadequate or inconsistent?	no	yes

the recognition test that was found in a pretest. Pretests also indicated that documenting the evidence using a checklist required less time than writing notes. In order to equalize the overall time spent on the first phase tasks; we took the following steps. First, in order to shorten the time spent by notes participants, instead of formal memoranda, we had them "jot down some brief notes that you could use later on to document the factors associated with management fraud," following the approach that Pincus used (1989). Second, we included a filler task for the Checklist condition that required the participants to perform a short, unrelated task using nine financial ratios to predict financial distress.

After documenting the experimental company's evidence, reviewing the second company, and rating their confidence, both groups saw a screen that said they had completed the first phase of the study and could not complete the second phase until the next day. The program included a programmed delay and would not allow the participants to begin the memory test until the following day. The delay presented a more realistic audit situation, in that time often passes before evidence considered at one point in an audit and is used to make decisions at another point; plus, the delay ensures that participants relied on their LTM.

When the participants reinserted the disk to start Phase II, the initial instructions for the recognition test prompted all participants to indicate the name of the hypothetical client about whom they were to answer questions.[6] All participants indicated the name for the proper hypothetical client. Regardless of MODE, the recognition test consisted of 24 statements (drawn from the questions in Exhibit 1) relevant to the presence or absence of fraud, and six filler items about client facts not associated with the likelihood of fraud. Each screen of the recognition test included the instructions that "If the above statement, taken as a whole, is supported by the evidence, click the 'yes' button. Otherwise, click the 'no' button."[7]

The final step in the experimental task was to "provide your overall assessment of the risk of management fraud for this client." The assessment was provided by clicking one of eleven buttons, labeled from "0%" to "100%." When the participants had completed the study, they were instructed to return the disk to the regional HR director.

Participants

The participants in this study were auditors from an international accounting firm who reported spending an average of 92% of their time conducting audits. The number of years reported ranged from 1 to six, with an average of 3.1 years. The distribution of levels within the firm was four staff auditors, 56

seniors, and six managers.[8] Comparisons between the levels of MODE show that there are no differences due to job title ($\chi^2 = 0.43$, $p < 0.98$), years experience ($F = 0.03$, $p < 0.87$), percent of time spent on audits ($F = 0.20$, $p < 0.68$), or area of specialty $\chi^2 = 0.001$, $p < 0.98$). Correlating each of these variables with recognition accuracy, we find that neither percentage of time spent conducting audits ($r = -0.15$, $p < 0.23$), number of years ($r = 0.19$, $p < 0.12$), nor job title ($r = 0.11$, $p < .36$) is correlated significantly our dependent variable.

ANALYSIS AND RESULTS

Preliminary Analysis

As a preliminary analysis, we examine the information processing of the participants in our study in three ways: (1) the time spent viewing the evidence; (2) the number of visits to the evidence screens; and (3) the time spent at the documentation screens. The time viewing the evidence related to the second client for which no documentation requirements existed is also measured. These four measures are analyzed using a single factor MANOVA with MODE as the independent variable (Wilk's Lambda = 0.65, $F = 8.2$, $p < 0.00$). Table 1 shows the means for each of the four dependent variables. The mean number of minutes spent viewing the evidence by participants in the notes group (15.2) is significantly greater ($F = 8.84$, $p < 0.01$) than the mean number of minutes spent by participants in the checklist group (11.3). Likewise, notes participants made an average of 5.8 visits to the evidence screens, compared to only 3.3 visits made by checklist participants ($F = 9.0$, $p < 0.00$). In contrast, in the absence of a requirement to document their understanding, both groups spent approximately equal amounts of time viewing the evidence for the second client ($F = 0.07$, $p < 0.79$).

Differences in information processing continued into the documentation screens, as notes participants spent a mean 16.9 minutes documenting whereas checklist participants spent a mean of 5.2 minutes. The latter result is impressive given that notes participants documented an average of 5.8 indicators, whereas checklist participants documented by responding to 12 questions. In total, the time that checklist participants spent viewing evidence plus documenting their understanding (16.5 minutes) is about 1.4 minutes per item documented. In contrast, the total time that notes participants spent (32.1 minutes) is about six minutes per item documented. These differences are so great as to preclude typing as an explanation, especially in light of the fact that the notes group was instructed to provide only brief notes. These findings suggest that documentation mode affected the amount of information processing.

Table 1. Mean Values for Four Measures of Information Processing During Evidence Documentation.

	Documented Client			Non-Documented Client
	Minutes Viewing	Number of Visits to Evidence	Minutes Documenting	Minutes Viewing Evidence
Checklist	11.3	3.3	5.2	7.5
Notes	15.2	5.8	16.9	7.9
F(1,64)	8.84	9.00	26.69	0.07
P-value <	0.00	0.00	0.00	0.79

Before proceeding with the analysis, it is necessary to define two additional terms, DOC and "as documented." The within-subjects variable DOC measures which items each participant documented. The twelve items appearing on the checklist were simply counted as having been documented by checklist participants. For notes participants, independent raters with significant audit experience examined documentation. The raters coded items as having been documented if they considered the participant's documentation addressed an item on the *recognition test.*[9] This measure separates out the recognition test items that would have been previously subjected to the special processes of being documented in notes. Furthermore, some tests are based on counts of documented items "as they were documented." The evidence covered in the documentation prompts (checklists and notes instructions) included three positive and nine negative indicators of fraud. Our measure of "as documented" is based on meaning. The meaning of what participants encoded into their memories at the time of documentation is ascribed by examination of the documented items. A checklist participant who responded "yes" to a checklist item, even though the actual fraud indicator in the evidence was negative, was counted as having documented the item as indicating the existence of fraud.[10]

Based on the findings of Pincus (1989), notes users can be expected to include more positive than negative fraud indicators. Examining the items documented by those using notes reveals that 74.4% of the documented items were positive fraud indicators. Those using checklists recorded an average of 4.2 (35% of the twelve documented items) as positive indicators of fraud. Figure 2 shows the distributions of the proportions of the positive indicators as documented. In order to test these distributions differ, it was necessary to collapse the scale into four categories (zero to 20%, 21%–50%, 51%–80% and 81%–100%) due to small cell sizes. Testing the collapsed scale shows that the two distributions differ ($\chi^2 = 33.48$, d.f. = 3, $p < 0.00$). In addition to the differences reported above, it should be noted that the two distributions overlap extensively, and

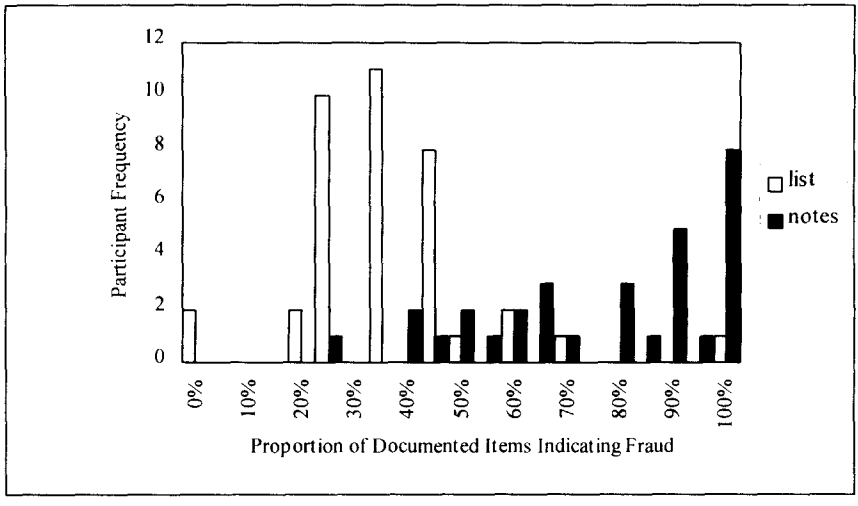


Fig. 2. Proportion Of Documented Items That Indicate Fraud By Documentation Condition.

These distributions are significantly different ($\chi^2 = 33.48$, d.f. = 3, $p < 0.00$).

that the notes distribution is flatter than the checklist distribution. As we argue later, it is important to understand that, while there are overall differences in the proportion of fraud indicators as documented by the two groups, these differences do not hold for all participants in either condition.

Effect of MODE on Memory Retrieval

Hypotheses 1, 2, and 3 relate to the effects of MODE on memory retrieval, and are tested using participants' responses to the 24 target items on a 30-item recognition test.[11] Of the target items on the recognition test, 12 were on the checklist (and prompted by the instructions for the notes group), and 12 were not.[12] Recognition test items were counterbalanced into sets of six, where all four possible combinations of "supported/not supported by the evidence" and "correct answer is yes/no" were represented. Recognition accuracy was the proportion of items on the test recorded accurately as positive or negative indicators, according to whether the attribute was explicitly described as present or absent for the client.

Table 2. Logistic Regression.

Variable	DF	Parameter Estimate	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	1	−1.25	117.18	0.00	
DOC	1	−0.36	4.23	0.04	−1.44
MODE	1	0.33	5.11	0.02	1.39
DOC × MODE	1	−0.05	0.03	0.85	−0.13

Dependent variable: Recognition rates.
Independent variables: whether or not an item was documented (DOC) and whether checklists or notes were used to document (MODE).

Recognition of Documented vs. Undocumented Evidence

Hypothesis 1 basically predicts that the process of documenting evidence will make documented items more memorable, so recognition for documented items will be higher than for undocumented items. Initially, a logistic regression, shown in Table 2, was conducted using the probability of recognition as the dependent variable; and DOC, MODE and their interaction as independent variables. Overall, the model fitted was significant $\chi^2 = 20.38$, d.f. = 3, $p < 0.00$), and it shows that both DOC and MODE were significant ($p < 0.04$ and $p < 0.02$, respectively). The significance of DOC in the overall analysis indicates an effect for the act of documenting. The frequency counts and percentages for correct and incorrect recognition are shown in Table 3. The comparison of documented vs. undocumented items shows that documented items were correctly recognized 82.2% of the time, while undocumented items were correctly recognized 74.2% of the time. Thus, Hypothesis 1 is supported by our results.

Comparisons Between Checklist and Notes

Hypothesis 2 predicts that, when only documented items are compared across documentation modes, the recognition rates for those writing notes will be higher than for those using checklists. Hypothesis 3 predicts that when undocumented items are compared across documentation modes, those using checklists will have a higher recognition rate than those writing notes. Taken together, hypotheses 2 and 3 predict an interaction between documentation mode and whether or not an item was documented. The interaction between DOC and MODE was not significant in the logistic regression and, therefore, the predicted interaction was not found; hence, neither Hypotheses 2 nor 3 were supported.[13] Instead, we found an overall effect for MODE. The recognition

Table 3. Recognition Rates.

MODE	DOC	Correctly Recognized	Incorrectly Recognized
Checklist	*documented*	360 (83.33)	72 (16.67)
	not documented	336 (77.78)	96 (22.22)
Notes	*documented*	128 (79.01)	34 (20.99)
	not documented	416 (71.43)	166 (28.57)
Overall	*documented*	488 (82.15)	106 (18.85)
	not documented	752 (74.16)	262 (25.84)

Recognition rates by documentation mode (checklist or written documentation) and whether the item was documented or not documented by the subject, and tests of differences between modes. Percentages of items either correctly or incorrectly recognized for each Mode by DOC combination are shown in parentheses.

rate for the notes condition was 73.1% and the rate for the checklist condition was 80.6%. While the evidence is not conclusive, we speculate that the cognitive overhead of writing notes is sufficiently taxing on the auditors that, whether or not an item is documented, it is less retrievable later for notes users.

Effects of MODE on Fraud Assessments

Hypothesis 4 begins to explore the nature of differences in judgment processes possibly caused by MODE. We tested for the effect of documentation mode on fraud assessments by comparing the mean judgments between the two documentation conditions and found no difference (Checklist = 4.03; Notes = 4.17; $F = 0.05$, $p < 0.82$). The variances in judgments across the two modes also were not different ($F = 1.1$, $p < 0.77$). Finding no effect for documentation mode on the fraud risk assessments is inconsistent with both the specific expectations

we developed based on our characterization of the cognitive processes evoked by alternative documentation modes and Pincus' (1989) fraud case. However, the finding of no difference in fraud assessment by MODE is consistent with the findings of Pincus (1989) in the case where no fraud existed. Having memory data for our participants allows us to explore the possible relationships among MODE, auditors' memories for the evidence, and subsequent judgments in greater detail in an effort to understand this lack of a direct link between MODE and judgment. Below we present several tests to shed light on this issue.

For the notes condition, the recognition rates for positive fraud indicators (80.1%) is higher than the recognition rate for negative indicators (73.1%). The difference between these recognition rates is significant ($\chi^2 = 3.62$, $p < 0.06$).[14] Then, as shown in Table 4, accuracy for each indicator-type (whether an item indicated the presence of fraud or not) was correlated with judgment. Those who used checklists had a significant, positive correlation ($r = 0.33$, $p < 0.05$) between their judgments and their recognition accuracy for positive indicators, and no relationship for negative indicators. For those who used notes, the correlation between their judgments and their recognition accuracy for negative indicators is significant and negative ($r = -0.38$, $p < 0.04$). The correlation between their assessments and their recognition accuracy for positive fraud indicators was not different from zero. This helps isolate the differences between documenting with checklists and using notes, and helps explain the link between MODE and judgment. While variation in what individual auditors remember appears to obscure the impact of MODE on judgment, this analysis supports the assertion that differences in what auditors retrieve from memory affects their judgment. The link between documentation mode and what gets encoded or is retrievable from memory has already been established.

Table 4. Spearman Correlation Between The Participants' Overall Risk Assessment And Accuracy On Recognition Test.

MODE	Fraud Indication	Spearman Correlation
Checklist ($n = 36$)	positive	0.33 (0.05)
	negative	–0.15 (0.39)
Notes ($n = 31$)	positive	–0.14 (0.21)
	negative	–0.38 (0.04)

The probability of the correlation being zero is shown in parentheses.

Finally, pursuing this result further, we dichotomized judgments into two categories. The "lower" classification includes all risk judgments of 0.6 and below, and the "higher" classification includes all the others.[15] Table 5 shows the mean recognition accuracy rates for both positive and negative indicators in each combination of MODE and risk category. The two means that stand out in this analysis are in the higher judgment category. Those who used checklists and have the highest recognition rate for positive fraud indicators of any condition (92%) rated the risk of fraud higher. In contrast, those who wrote notes and have the lowest recognition rate for negative indicators of any condition (62%) also rated the risk of fraud higher. Thus, it appears that when members of the checklist group accurately remembered more positive fraud indicators, they rated the risk as higher. In contrast, when members of the notes group inaccurately retrieved negative fraud indicators as positive, they rated the risk as higher.

DISCUSSION AND IMPLICATIONS

Before discussing the results and suggesting related implications, it is important to point out some of the study's limitations and strengths. As with all studies conducted in the laboratory, caution is advised when generalizing results to the field based on a single study. In order to achieve the high level of internal validity necessary to examine the theoretical issues, i.e. the link between documentation mode and memory, some features of the audit environment were not considered. Such features include, but are not limited to, time pressure, accountability, the personal relationship of the auditor to his or her colleagues and

Table 5. Mean Recognition And Sample Sizes Rates.

		Fraud Risk Assessment	
		Low	High
Checklist	positive $n = 30$	0.78	0.92
	negative $n = 6$	0.81	0.81
Notes	positive $n = 24$	0.81	0.76
	negative $n = 8$	0.76	0.62

client personnel, and the personal or professional pressures to perform. Inclusion of all these features is beyond the scope of any single study. A strength of the study is the added control provided by its experimental design. This includes prompting the subjects in both conditions to consider the same items, thus ensuring that any differences found in the memory test were due to documentation mode rather than what was considered. We also designed the study to negate differences in the time spent by each group. Together with the between subjects' design, these features provide increased internal validity for examining the theoretical issues at hand.

Calls for improved external auditing (Levitt, 1998) and better understanding of the role of documentation in the audit process (POB, 2000) underlie the research into evidence documentation found in this study. We examined the relationships among auditors' documentation mode, memories for the documented evidence, and judgments. Two groups of auditors, a note group that provided written documentation and a checklist group that used a prepared checklist, performed an experimental task that required them to evaluate and document fraud-related evidence and assess the potential for management fraud. Checklists are completed faster, require less time reviewing evidence and are more comprehensive. We also found that documentation mode affected the nature of the fraud-related evidence that an auditor documented. Consistent with prior research (Pincus, 1989), those who used notes to document tended to include a higher proportion of positive fraud indicators. They actively select positive fraud indicators (or presumably positive indicators for whatever their judgment task happens to be) and fail to select negative indicators. This difference in the nature of the items documented between checklists and written notes could have an important effect on the review process, since the contribution of the review process is likely to be reduced if what is documented is not representative of the evidence that was evaluated. These differences argue for use of checklists to document evidence in situations where the evidence items can be specified before the fact so that comprehensive list of items to be considered can be constructed.

The major contribution of this study was the inclusion of memory measures based on the participants' recognition of the evidence they reviewed. The comparison of documented vs. undocumented items shows that documented items were correctly recognized significantly more often than undocumented items. We conclude that regardless of the mode used, the process of documenting makes evidence more memorable. We also found that the recognition rate for the notes condition was significantly lower than for the checklist condition. The cognitive processing demands of writing notes appears to be sufficiently taxing on the auditors' cognitive abilities that, whether or not an item is

documented, it is less retrievable later for notes users. While the documentation process inherent in use of notes forces auditors to actively select and evaluate items of evidence, the process of constructing the written documentation appears to be sufficiently demanding so as to offset the benefits of deeper cognitive processing.

Tests for an effect of documentation mode on fraud found no overall direct effect of documentation mode on subsequent judgments. The variances in judgments also were not different across the two modes. Finding no difference in fraud assessment across documentation mode is consistent with Pincus' (1989) case where no fraud existed. However, our findings were inconsistent with the expectations we developed and Pincus' (1989) fraud case. We suspect that a difference in experimental design explains why Pincus find some differences in fraud judgments between documentation conditions. Meaning, Pincus based her case in an actual fraud case and presented the case in two forms – with and without fraud. Our case was patterned after a case used in audit training by a major public accounting firm, not actual facts, which may be considered a limitation of our study. While we do not have external validation of the possibility of fraud in our case, there were only three positive indicators of fraud in our narrative, which may not be sufficient to indicate fraud. If this assumption holds, our results are not in conflict with those of Pincus because she also found no difference due to documentation mode in her no-fraud case.

Using the memory data for our participants allows us to explore the possible relationships among documentation mode, auditors' memories for the evidence, and fraud assessments in greater depth than prior studies. Ex post analyses compared recognition rates between positive and negative fraud indicators for the two documentation conditions and found that, among the four combinations of indicator-type and documentation mode, checklist users' recognition for positive indicators was highest and notes documenters' recognition rate for negative indicators was lowest. In addition, checklist users' assessments were significantly positively correlated with recognition rates for positive items, while notes users' assessments were significantly negatively correlated with recognition rates for negative items. Thus, in trying to explain the link between documentation mode and judgment, we found that use of a checklist generally leads to better recognition for positive fraud indicators and that the positive-indicator recognition rate was positively correlated with checklist users' risk assessments. Notes users, on the other hand, reflected lowest recognition rates for negative indicators and the negative-indicator recognition rate was significantly negatively correlated with risk assessments. Thus, checklist use seems to lead to better recognition of positive items and, as their memory for positive indicators increased, so did their perceived likelihood of fraud. Notes

takers' judgments seem to be driven by what they could not remember; their recognition of negative indicators on average was low and, as it got worse, their assessment of the likelihood of fraud were higher. Thus, we offer some evidence that the nature and contents of an auditor's memory, which was conditioned by documentation mode, mediated between mode and judgment.

The results of this study have important implications for auditing research and practice. First, the mediating role of memory helps explain why neither our research nor that of Pincus (1989) found a systematic relationship between the way evidence was documented and the judgments that were made. Our results show that the nature and content of an auditor's memory is conditioned by documentation mode, and that memories mediate between mode and judgment. This suggests that audit firms, and possibly standard setters, should recognize that evidence documentation by either method can influence what an auditor remembers. Checklists must be comprehensive enough that auditors who use them will consider not only positive indicators, but also indicators that highlight countervailing factors. If less structured documentation is used, methods need to be devised that make counter indicators more memorable. Both of these measures would insure that the risk of fraud is not over-estimated leading to an inefficient audit.

Even without the requirements of professional standards, audit firms need documentation. How the process of documentation evolves will be a function of many factors, not the least of which are litigation and technology. Further understanding of the documentation process, as provided in this study, will serve as valuable input to improving audit practice. Among the possible avenues for further research are: (1) a more explicit study of the role of experience in documenting fraud assessments; (2) examining the potential differences between internal and external auditors in fraud detentions; and (3) the idea of a confirmation bias in auditors' evidence documentation.

NOTES

1. In his address, Chairman Levitt proposed a public oversight board to "review the way audits are performed and assess the impact of recent trends on the public interest." That board was established and issued its final report in August 2000. The report notes that, in general, auditing standards leave the extent of documentation to the judgment of the auditor. The panel used SAS No. 82 as an example of the type of documentation that it recommends. SAS No. 82 specifies certain aspects of what should be documented, but does not address the means by which that assessment should be documented. With regard to documentation of internal controls it found that documentation "takes different forms, including firm-specific checklists and preparation of detailed descriptions of the entities policies and procedures" (POB, 2000, p. 27).

2. The *t*-statistic for difference in mean values had a probability of 0.069.

3. According to Shelton et al. (1999) even though firms encourage or require fraud risk assessments to be performed by a manager or partner, representatives from several firms indicated that in practice seniors might perform the review with concurrence by superiors. This coupled with the fact that seniors would be in training to perform this function means it is realistic to use them as participants in this study.

4. The instrument was created prior to the final release of SAS No. 82.

5. In order to provide every opportunity for the participants to calibrate their confidence ratings, a preliminary general assessment preceded the confidence probe on a separate screen. This probe asked for the participants' predicted performance in recognizing "general statements" supported by the evidence. Only the confidence ratings for detail and facts are analyzed as all target statements in the memory test were fact-oriented rather than general in nature.

6. This probe was to insure that participants were answering questions about the firm they had documented rather than the second, unrelated client.

7. While we did not anticipate any biases due to the frequency of responding "yes" or "no", we balanced our design so that exactly half the statements (whether in the documentation prompts or just on the recognition test) had "yes" as the correct response.

8. One participant did not report his/her job position.

9. The initial agreement between the raters was greater than 80% and the items used in the analysis were those that followed reconciliation of differences between the raters.

10. It is impossible to be certain from the documentation whether Notes participants considered items to be positive or negative indicators of fraud. Therefore, we assume that participants followed the professional guidelines and considered all and only the items we constructed as positive fraud indicators to be such. Our analysis reflects this assumption.

11. The items on the recognition test were not taken verbatim from the case text, but were written to convey the meaning without the context of the narrative. For example, the evidence narrative said "The previous firm did not report knowledge of any significant problems on the audit or relations with management." The related recognition test item said, "The previous auditor noted no significant audit exceptions, and no reason to believe that fraud had ever occurred."

12. Pretests were used to minimize the possibility that participants would not identify items other than those on the 24-item list as being important. Our participants in the notes condition, however, did document some items in addition to those on the recognition test. These items were very limited in number and included things like references to the internal audit staff. We have not included a summary since there were so few items and their relation to judgment cannot be tested.

13. Lack of a significant interaction implies that that mean values that would be tested in *H2* (i.e. only documented items compared across documentation modes) and *H3* (i.e. only undocumented compared across documentation modes) are not significantly different

14. The recognition rate for positive and negative indicators was identical, 80.6%.

15. We do not believe the relationship between memory for fraud items and judgment is linear. We conducted additional analyses using cut-off points other than 0.6 to dichotomize the participants' likelihood assessments. The results of those analyses were substantially the same, though weaker.

ACKNOWLEDGMENTS

This research was supported by the General Research Fund and the E&Y Center for Audit Research both at the University of Kansas. We thank Peter Gillett, Keith Harrison and Margaret Reed for their help in data collection and analysis. We thank the participants at workshops at the Universities of Colorado, Kansas, Kentucky, Missouri, North Texas, Portland State, and Utah. We particularly thank Jim Frederickson, Marlys Lipe, Alan Mayper and Bob Ramsay. The authors are listed in arbitrary order; each made equal research contributions.

REFERENCES

American Institute of Certified Public Accountants (1997). *Professional Standards* (Vol. 1). New York: AICPA.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671–684.

Choo, F., & Trotman, K. (1991). The relationship between knowledge structure and judgments for experienced and inexperienced auditors. *The Accounting Review*, (3), 464–485.

Eining, M., Jones, D., & Loebbecke, J. (1997). Reliance on decision aids: an examination of auditors assessment of management fraud. *Auditing: A Journal of Theory and Practice*, (Fall), 1–19.

Levitt, A. (1998). The numbers game. remarks by Chairman Arthur Levitt delivered at the NYU Center for Law and Business, New York, NY, September 28, 1998.

Moeckel, C. (1990). The effect of experience on auditors' memory errors. *Journal of Accounting Research*, (Autumn), 368–387.

Moeckel, C., & Plumlee, R. D. (1989). Auditors' confidence in recognition of audit evidence , *The Accounting Review*, (October), 653–666.

Libby, R., & Trotman, K. (1993). The review process as a control for differential recall of evidence in auditor judgments. *Accounting, Organizations and Society*, (6), 559–575.

Loebbecke, J., & Willingham, J. (1988). Review of sec accounting and auditing enforcement releases. Unpublished working paper, University of Utah.

Moser, D. (1989). The effect of output interference, availability and accounting information on investors' predictive judgment. *The Accounting Review*, (3), 433–448.

Nelson, M., Libby, R., & Bonner, S. (1995). Knowledge structure and the estimation of conditional probabilities in audit planning. *The Accounting Review*, (1), 27–47.

Public Oversight Board (2000). The panel on audit effectiveness report and recommendations. Stamford, Connecticut, The Public Oversight Board.

Pincus, K.. V. (1989). The efficacy of a red flags questionnaire for assessing the possibility of fraud. *Accounting Organizations and Society, 14*, 153–163.

Purvis, S. E. C. (1989). The effect of audit documentation format on data collection. *Accounting Organizations and Society*, 14: 551–563.

Ramos, M. J., & Lyons, A. M. (1997). Considering fraud in a financial statement audit: practical guidance for applying SAS no. 82. AICPA, New York.

Ricchiute, D. N. (1997). Effects of judgment on memory: experiments in recognition bias and process dissociation in a professional judgment task. *Organizational Behavior and Human Decision Processes, 70*(1), 27–39.

Shelton, S., Whittington, R., & Landsittel, D. (1999). Fraud risk assessment: an analysis of practices of auditing firms. Working paper DePaul University.

Zimbelman, M. (1997). The effects of sas no. 82 on auditors" attention to fraud risk factors and audit planning decisions. *Journal of Accounting Research, 28*(Supplement), 75–97.

OUTCOME INFORMATION AND THE EVALUATION OF AUDITOR PERFORMANCE: THE ROLE OF EVIDENCE RECALL, INTERPRETATION AND WEIGHTING

John G. Wermert

ABSTRACT

While prior research has found that outcome information affects jurors' evaluations of auditor performance, little is known about the underlying cognitive processes that give rise to this outcome information effect in auditor liability settings. This study extends previous research in accounting and psychology by investigating whether jurors' recall, interpretation and weighting of trial evidence mediates the relationship between outcome information and the evaluation of the auditors' performance. The research question was examined using an experiment with 71 surrogate jurors. Path analytic modeling reveals that the effect of outcome information on the evaluation of the auditors' performance is mediated by the jurors' interpretation and weighting of the trial evidence, but that recall had no impact on the performance evaluations of the auditors.

Advances in Accounting Behavioral Research, Volume 5, pages 77–113.

ISBN: 0-7623-0953-9

INTRODUCTION

During the 1990s the auditing profession faced a litigation crisis that saw the largest audit firms in the United States (then the "Big 6") spending inordinate sums on litigation. On an annual basis, such expenditures by the Big 6 collectively reached as much as $1 billion, which was approximately 20% of the Big 6 firms' annual audit and accounting revenue (Mednick, 1996). Even after factoring in insurance recoveries, litigation costs still amounted to an astounding 11.9% of revenues. Furthermore, this trend of large payments by CPA firms seems to continue unabated. As Berton (1999a) notes, CPA firms are "still coughing up the dough." This crisis was not limited to the Big 6 as regional and local firms were likewise affected (*CPA Journal*, 1990).[1]

In response to the rash of litigation in the 80s and 90s, accountants sought and won some important judicial and tort reforms, such as The Private Securities Litigation Reform Act of 1995 (Cloyd et al., 1998). Despite such reforms, the litigation crisis is far from over (Cloyd et al., 1998), since predicting the future course of litigation against auditors is difficult due to the evolutionary nature of the law and because judges/courts may attempt to circumvent national reforms (Silicano, 1997). Plaintiff's attorneys will likely shift much of their litigation against auditors to more favorable state venues in an attempt to circumvent reforms at the national level (Palmrose, 1997a, p. 356). In addition, Berton (1999b) reports that affiliates of major accounting firms throughout the world are increasingly becoming targets of litigation, which is likely to boost the costs of auditing in the United States.

Given the substantial effect of litigation on the auditing profession in the past, as well as its likely continued effect in the future, understanding the factors that affect the ex post evaluation of auditor performance is critical (Kinney, 1993, 1994; Palmrose, 1997a). One such factor that has been investigated in prior accounting and auditing research is the availability of outcome information (J. Anderson et al., 1997; Kadous, 2000; Lowe & Reckers, 1994). This research has found that information regarding ex post audit outcomes that is available to the evaluator (e.g. judge, juror, or regulator) often has a very significant affect on the evaluation of the auditor, even in settings where using such information is clearly not appropriate.

This study extends this prior research by exploring the cognitive factors that give rise to the outcome information effect in auditor liability settings. This extension is important because prior research in psychology has shown that outcome effects are context specific (Hawkins & Hastie, 1990), which underscores the importance of studying outcome effects in the particular context of interest. Furthermore, understanding the source of outcome effects in auditor

liability settings should allow auditors to work proactively to reduce their exposure to outcome effects ex ante, or to select the most appropriate strategies in court ex post to mitigate the *specific* cognitive factors producing the effect. Accordingly, this study hypothesizes that outcome knowledge influences the jurors' recall, interpretation and weighting of trial evidence and that these processes mediate the relationship between outcome information and performance evaluation.

A MODEL OF JUROR DECISION MAKING

This study uses a general model of juror decision making based upon information integration theory (N. Anderson, 1981, 1974) to develop hypotheses regarding how outcome information affects the ex post evaluation of auditor performance. The general model represents a juror's decision regarding the likelihood of "guilt" as the outcome of an information integration process described in terms of an algebraic combination rule (Hastie, 1993).[2] The juror combines the evidence presented at trial to form a subjective estimate of the likelihood of guilt, which is then compared to the appropriate standard of proof (i.e. reasonable doubt or preponderance of the evidence). If the likelihood exceeds the standard of proof, the juror returns a verdict in favor of the plaintiff or prosecution. Otherwise, the judgment is "for" the defendant. Information integration models have been found at work in many different decision-making tasks including juror decision-making (Hastie, 1993; Kaplan & Miller, 1978; Ostrom et al., 1978).

According to the model, a juror's subjective likelihood of guilt (J) is a weighted average of their initial opinion and the information obtained during the trial (Hastie, 1993):

$$J = \frac{s_0 w_0 + \sum_{i=1}^{k} s_i w_i}{w_0 + \sum_{i=1}^{k} w_i} \tag{1}$$

J can range from 0 to 1 with 0 indicating an absolute belief of innocence and 1 indicating an absolute belief of guilt. s_i represents the subjective value of guilt conveyed by a particular piece of evidence, again ranging from 0 (total innocence) to 1 (total guilt), with s_0 indicating the juror's initial belief. w_i indicates the weight or importance assigned to each of the k pieces of evidence, or the initial opinion (w_0). The model explicitly assumes that jurors are likely to have preconceived opinions or attitudes (s_0). However, prior attitudes will not affect jurors' final decisions if the jurors set aside their prior beliefs and assign a weight

of 0 to the prior attitude. Such a weight is implied by the instruction to the jurors to base their decision only upon the evidence presented during the trial and to disregard any preconceived opinions or information that they may have heard about the case from other sources (e.g. newspapers) before the trial began.

The process by which the juror calculates J and returns a verdict includes several distinct steps: evidence recall, interpretation (or valuation), weighting, and integration. First, the juror must select the admissible evidence from all of the inputs encoded from the trial (Pennington & Hastie, 1981). The judge's instructions often specify what may or may not be considered evidence. For example, the fact that someone is indicted or named as a defendant in a lawsuit should not be considered evidence nor should the attorneys' opening statements, closing arguments, or questions asked by the judge and attorneys. Additionally, other inadmissible evidence may have been presented at trial and the judge may have instructed the jury to disregard this evidence.

Following the selection of evidence, the juror interprets the evidence and converts each item of evidence to a scale value s_i. Anderson (1981, p. 5) refers to this operation as the process by which information is extracted from the physical stimulus using some dimension of judgment (e.g. probability of guilt) set out in the task instructions. Thus, in terms of the juror decision-making model this interpretation operation denotes the chain of processes that lead from the evidence presented at trial to the subjective probability of guilt conveyed by the piece of evidence.

Following the interpretation process, jurors must assign weights to each piece of evidence and their initial belief of guilt or innocence. Ideally, jurors should assign a weight of 0 to their initial belief since under American law defendants are presumed innocent until proven guilty. Weights assigned to the evidence presented during a trial will depend upon many different factors including (but not limited to) evidence credibility, relevance and importance.

After the scale values (s_i) and weights (w_i) have been determined for each item of evidence, the information is integrated by weighting the scale values by the weights and summing the information into a judgment (J) of the likelihood of guilt using Eq. (1). The judgment is compared to the standard of proof required to convict. A verdict of guilty is returned if the standard of proof is met or exceeded.

ROLE OF OUTCOME INFORMATION IN THE MODEL AND DEVELOPMENT OF HYPOTHESES

The previous section presented a general model of juror decision-making. This section applies the general model to lawsuits against auditors for failure to

detect fraudulent financial reporting by management. Hypotheses are developed regarding the role of outcome information as well as the cognitive factors that give rise to the outcome information effect. The context used in this study, the failure of auditors to detect management fraud, was chosen as the backdrop for the current study for several reasons. First, management fraud detection is an area that is likely to be very susceptible to outcome information effects as red flags may be far easier to identify ex post than ex ante (Buchman, 1985). Second, almost half of all lawsuits against auditors of bankrupt companies include allegations of undetected management fraud (Palmrose, 1997b). Third, Congress and the SEC have both been highly critical of auditors' performance in detecting management fraud (Hearings, 1977, 1989; General Accounting Office, 1989; MacDonald, 1999; Treadway Commission, 1987), and this study will be useful in determining the extent that outcome effects may contribute to such perceived deficiencies. Finally, this is an area that is likely to see increased regulatory attention since the SEC has publicly stated that it plans to more aggressively prosecute fraudulent financial reporting in the future (MacDonald, 1999; Schroeder, 2001).

Prior research on outcome information effects has shown that bad outcomes lead to unfavorable evaluations of performance and good outcomes lead to favorable evaluations (Anderson et al., 1997; Baron & Hershey, 1988; Brown & Solomon, 1987, 1993; Lipe, 1993; Lowe & Reckers, 1994; Kadous, 2000). Therefore, in the context of a lawsuit against an auditor for failure to detect management fraud, a juror who knows that management fraud (the "bad outcome") was discovered after an audit was completed should be more likely to believe that an auditor breached the standard of care (Buchman, 1985). Alternatively, a juror may be more likely to conclude that the standard of care was adequate if a good outcome occurs than if either no outcome information is available or a bad outcome occurs. This suggests the following hypothesis, which is a replication of prior work in a management fraud context:

> H_1: Juror evaluations of auditor performance will be affected by outcome knowledge with evaluations being adversely affected by bad outcomes (occurrence of fraud) and favorably affected by good outcomes (the non-occurrence of fraud).

As stated in the introduction, the main focus of this paper is on understanding the cognitive mechanisms that give rise to the outcome information effect in auditor liability settings. This is accomplished in this study by examining the role of outcome information in the relevant subtasks of the judgment process: evidence recall, evidence interpretation and evidence weighting.[3]

Evidence Recall

When evaluating an auditor, the jurors' first task is to recall the admissible evidence from all of the inputs encoded from the trial (Pennington & Hastie, 1981). However, as suggested by Slovic and Fischhoff (1977), outcome information may affect evidence recall. The recall of evidence is a deliberate task and jurors may use outcome information as a cue to retrieve evidence. In lawsuits against auditors, if jurors use the reported outcome (e.g. management fraud) as a cue to recall information, the jurors may retrieve more evidence that is consistent with the reported outcome and less evidence that is inconsistent with the reported outcome than they would have otherwise retrieved in foresight. Thus, outcome information may result in "biased retrieval." This is similar to prior research that found that subjects could recall more facts that supported decisions they had made than facts that contradicted their decisions (Brown & Solomon, 1993; Dellarosa & Bourne, 1984). In addition, the type of evidence recalled, either outcome consistent or inconsistent, should directly influence the performance evaluation of the auditor by a juror. Recall of evidence that is consistent with fraud should adversely affect the evaluation of the auditor, and recall of evidence that is inconsistent with fraud should positively affect the auditor's evaluation. This leads to the second and third hypotheses:

H_2: Outcome information will affect the nature of the evidence recalled, with jurors recalling in hindsight more evidence that is consistent with the reported outcome than they would have otherwise recalled without the outcome information.

H_3: The nature of the evidence recalled (i.e. consistent or inconsistent with management fraud) will influence the performance evaluation of the auditor.

Evidence Interpretation

Following the recall of evidence, the juror must interpret the evidence. At this stage of the decision making process, the juror extracts information from the physical stimulus using some dimension of judgment (e.g. probability of guilt) set out in the task instructions (Anderson, 1981, p. 5). Outcome information is also expected to affect this valuation operation. In foresight, the proper interpretation or implication of evidence may seem ambiguous. However, after the outcome is known, the evidence may seem to clearly imply the outcome (Slovic

& Fischhoff, 1977). Fischhoff (1975, p. 297) suggested that "upon receipt of outcome knowledge judges immediately assimilate it with what they already know about the event in question. In other words, the retrospective judge attempts to make sense, or a coherent whole, out of all that he knows about the event." In a litigation context, evidence that appeared rather innocuous ex ante may be interpreted by the jurors to be consistent with the outcome reported (Casper et al., 1988; Casper et al., 1989). This "retrospective sense making" is expected to directly affect the interpretation of evidence in auditor liability settings and this interpretation of the evidence is expected to have a direct effect on the performance evaluation of the auditor. This leads to hypotheses H_4 and H_5:

H_4: Outcome information will affect the interpretation of ambiguous evidence, with the evidence being interpreted in hindsight to be more consistent with the reported outcome than what would otherwise occur without the outcome information.

H_5: The manner in which evidence is interpreted will influence the performance evaluations of the auditor.

Evidence Weighting

In the last step of the decision process, the juror must combine the implications of the pieces of evidence to arrive at an overall decision. The most common method used to model this integration process is the use of an algebraic combination rule, where weights are applied to each of the pieces of evidence before they are combined to yield a final decision (Hawkins & Hastie, 1990; Anderson, 1981). Weights assigned to the evidence will depend upon many different factors including (but not limited to) evidence credibility, relevance and importance. Ceteris paribus, as each of these three characteristics increases, so does the weight assigned to the evidence.

Outcome information is expected to affect each of these characteristics. First, evidence that appears consistent with the outcome is likely to be considered more credible by a juror. For example, if two expert witnesses provide conflicting opinions about the appropriateness of a certain medical procedure that contributed to a patient's death, the expert who argues that the procedure was inappropriate may seem more credible due to the consistency between the outcome and his testimony even though in reality most doctors would have performed the same procedure under the circumstances (the testimony of the other expert witness). Jurors may not appreciate the probabilistic nature of some decision processes and

that good decisions ex ante can lead to bad outcomes resulting in inflated estimates of credibility for outcome consistent evidence. Second, evidence consistent with an outcome may seem more relevant than inconsistent evidence. Consistency between evidence and an outcome may imply that such evidence is relevant for the decision even though perhaps the relationship was coincidental or illusory. Jurors inexperienced in matters being judged may perceive causal relationships where none exist and decide mistakenly that irrelevant information was in fact relevant. Finally, outcome information may affect the perceived importance of evidence for determining guilt or innocence. Evidence that is consistent with the outcome will appear more important than evidence that is inconsistent since these factors seemingly led to or caused the outcome.

In the context of a juror evaluating an auditor's performance in detecting management fraud, the weighting of evidence is expected to be dependent upon the consistency between the evidence item and the reported outcome. Evidence consistent with the reported outcome is expected to be weighted more heavily in hindsight than in foresight. Evidence that is inconsistent with the reported outcome is expected to be weighted less heavily in hindsight than in foresight. Furthermore, the jurors' weighting of the evidence is expected to directly influence the performance evaluation of the auditor. This leads to the last two hypotheses:

H_6: Outcome information will affect the weighting of evidence, with evidence consistent (inconsistent) with the reported outcome being weighted more heavily (less heavily) in hindsight than in foresight.

H_7: The weighting of evidence by the performance evaluator will directly affect the evaluation of the auditor.

RESEARCH METHODOLOGY

The research hypotheses were tested in an experiment conducted on the campus of Indiana University. This section discusses the experimental design, participants, instrument and task, and the dependent, independent, and mediating variables used in the study.

Overview of Experimental Design

Subjects were randomly assigned to one of three different outcome conditions: no-outcome, bad-outcome, or good-outcome.[4] In the no-outcome group, subjects

were told that the company used in the case had been randomly selected from the client list of the auditor, and whether management fraud had occurred at the company during the year in question was not known. Subjects were informed that they should evaluate the auditor based upon the information provided in the case. Such an evaluation should be comparable to an evaluation performed in foresight.

Subjects in the bad-outcome group were informed that after the audit was completed allegations arose that management had intentionally overstated the value of inventory in its financial statements. A special fraud investigation concluded that management fraud had occurred. Subjects in the good-outcome group were also told of the allegations of fraud in the financial statements, however, they were told that the special fraud investigation concluded that no management fraud had occurred. While the good-outcome condition is somewhat artificial because litigation is not common when good outcomes occur, the results related to the good-outcome condition are still useful and interesting. Most audits do not result in audit failure, and accordingly, understanding how the performance of auditors is viewed when good outcomes occur is useful.

Participants

Seventy-one individuals, solicited through a variety of classes in the College of Arts and Sciences, participated in the experiment.[5] However, seven subjects were dropped from the experiment for various reasons: five subjects incorrectly answered the manipulation check at the end of the experiment, one subject indicated he had difficulty understanding the case, and two subjects missed multiple quiz questions used to determine if the subject was being attentive to the materials presented (one of these subjects also incorrectly answered the manipulation check). Thus, the results reported in this study are based upon the remaining sixty-four subjects. The seven subjects eliminated were almost evenly split amongst the three conditions included in the experiment. Furthermore, the results reported in this paper were not significantly affected by the elimination of the seven subjects previously mentioned.

Students from approximately 30 different majors from across the university were represented in the sample with education being the most heavily represented major (eleven subjects) and computer science being the second most frequent major (seven subjects). Approximately 61% of the sample was male and generally the sample was fairly young, as might be expected on a college campus. The mean age of the subject group was 21.1 years with a sample standard deviation of 2.3 years. Approximately 85% of the students were college juniors and seniors. Self-reported ratings of political beliefs were measured on

an 11-point scale ranging from very liberal (1) to very conservative (11). The mean response was 5.4 and the sample standard deviation was 2.5. Both very liberal and very conservative viewpoints were represented in the sample. All subjects were U.S. citizens.

Instrument and Task

An experiment was constructed using accepted guidelines for demonstrating the nonnormative use of outcome information (J. Anderson et al., 1997; Hershey & Baron, 1992, 1995). These guidelines require that: (1) subjects be provided with adequate information so that they will not be outcome dependent; (2) subjects are explicitly informed that they have all information that was available to the decision maker; and (3) subjects were explicitly directed to evaluate the decision maker as if they did not know the outcome.

The experiment required subjects to read through a case booklet and complete eight separate tasks. A summary of the eight tasks is included as Fig. 1, and an abbreviated form of the instrument is included as Appendix A.[6] The first task provided subjects with several pages of background information about financial statements and auditing. This information was designed to provide subjects with the background material that jurors would typically be provided during a trial and that was considered necessary to evaluate the performance of an auditor. After the subjects had read this information, they completed six questions to test their understanding of the material (Task 2).[7]

In the third task, subjects read the case that was used as the basis for the performance evaluation of the auditors. The case involved the audit of inventories of a consumer electronics retail chain by a large, international CPA firm. Most of the information in the case dealt with the auditors' observation of the physical inventories of the client company.[8] The case began with background information about the CPA firm followed by a manipulation of the outcome of the audit, the independent variable (details regarding the independent variable are discussed later in this section). This was followed by a description of the audit procedures performed by the CPA firm and the results of those tests. The case ended by stating that the auditors had concluded that inventory was fairly stated based upon the procedures performed. Approximately five pages of information was provided during the third task to ensure that the subjects were provided with adequate information so as not to be outcome dependent (Hershey & Baron, 1992, 1995; Anderson et al., 1997), but yet not impairing subjects with information overload (Casper et al., 1989, p. 297).

After the subjects had read the entire case, they were asked to assess the sufficiency of evidence gathered by the auditor (Task 4). Subjects were specifically

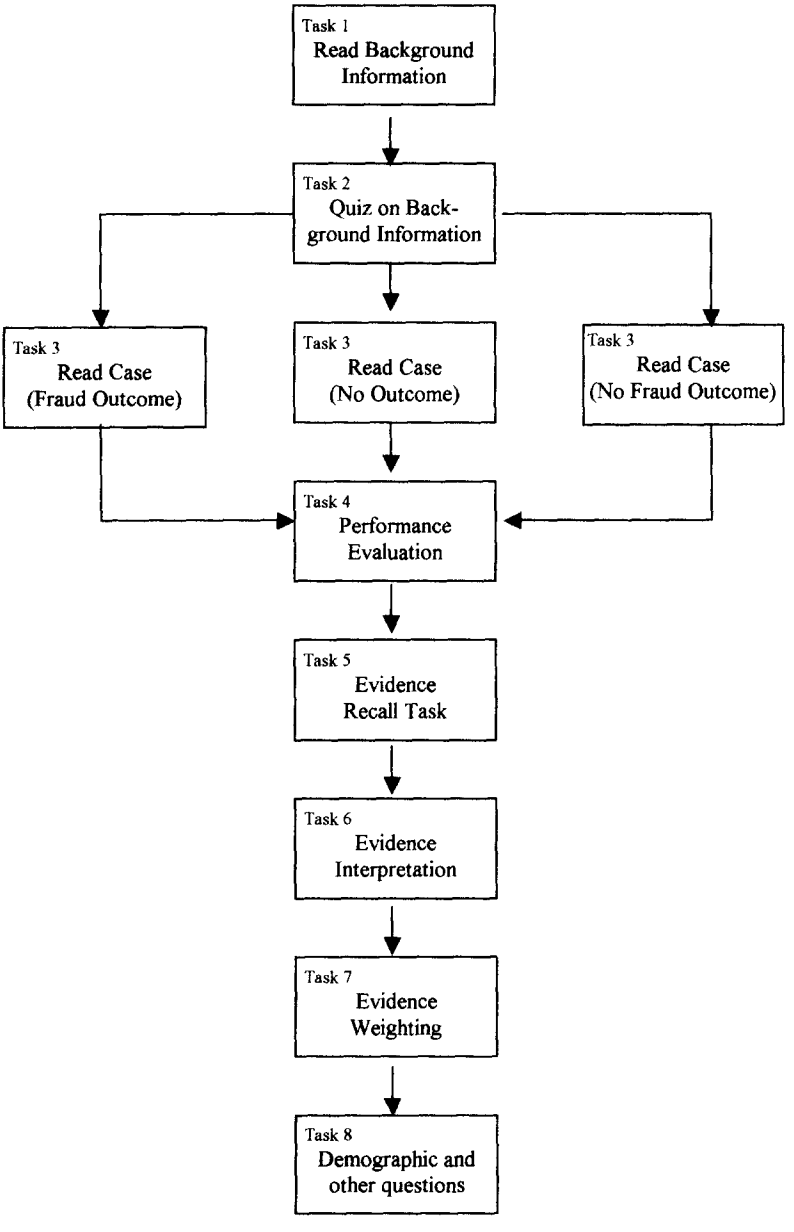


Fig. 1. Overview of Experimental Task.

told that they had been given all of the relevant information that was available to the auditors at the time of the audit and that they should base their evaluation of the auditor on this information. Furthermore, subjects who had received outcome information were specifically told that they should not allow their decisions to be affected by the outcome information. Next, subjects performed a recall task (Task 5), where they were asked to recall and list as many of the case facts as possible. Care was taken in designing the case scenario to include sufficient information (i.e. approximately 20 different pieces of evidence) to make the recall task meaningful. This was followed by an evidence interpretation task (Task 6) and an evidence weighting task (Task 7). In the final task, subjects provided demographic information, completed a manipulation check and indicated how understandable they had found the case (Task 8). Overall, subjects responded that they found the case very understandable, with the average response equal to 9.2 on an 11-point scale.

Compensation

Each subject received up to $15 of compensation for participating in the experiment. Subjects were informed of the pay scheme, which included both fixed and variable components, before beginning the experiment. First, subjects received $5 for participating in the experiment, regardless of their performance. This fixed component was included in the pay scheme to guarantee prospective subjects that they would have some earnings for participating in the experiment. Second, subjects earned up to $10 of additional compensation by correctly answering five questions on a post-experimental quiz, paid at a rate of $2 per question answered correctly. This $10 variable portion of the compensation was designed to encourage subjects to pay close attention to the case materials. Sixty-five of the subjects answered the post-experimental questions correctly thus earning the full $15. The other six subjects each missed one question earning $13 each.

Independent Variable

Outcome information, the sole independent variable, was manipulated in Task 3. Participants assigned to the no-outcome group were told that the company they were evaluating was selected randomly from the client list of the auditor, and no outcome information was provided. Subjects in the bad-outcome group were informed of the occurrence of management fraud at the beginning of task 3, the presentation of case facts. This is similar to trial settings where jurors would be informed of the outcome in the attorneys' opening

statements before the case information is presented. Subjects receiving the bad-outcome manipulation were provided the following paragraph:

> Quest's auditors did not detect any management fraud during the 1992 audit. However, after the 1992 audit was completed, allegations arose that Quest's management intentionally overstated the value of Quest's inventories in its 1992 financial statements. **A special fraud investigation concluded that management fraud had occurred. However, the fact that the auditors did not detect the management fraud does not necessarily imply that their performance was inadequate.** You should evaluate the auditors' performance using the information that was available to the auditors at the time they performed the audit. Please do not allow your decisions in this case to be affected by the conclusions of the special fraud investigation.

The good-outcome manipulation was identical to the bad-outcome manipulation paragraph except that the highlighted text read as follows:

> **A special fraud investigation concluded that no management fraud had occurred. However, the fact that no management fraud occurred does not necessarily imply that the auditor's performance was adequate.**

The instructions to the subjects to ignore the outcome information are important from a judicial standpoint. In such suits against auditors for failure to detect fraud, the defendant's actions should be judged based upon the circumstances existing at the time of the audit, and jurors should base their verdicts only on the evidence presented (Prosser & Keeton on Torts, 1985, Sec. 170; Sand et al., 1997). The courts have ruled consistently that an adverse outcome does not constitute or prove negligence (Staloch v. Holm, 1907, p. 264; Teig v. St. John's Hospital, 1963, p. 527).

Dependent Variable

The experiment included one dependent variable, performance evaluation, which was measured immediately after subjects had read all of the case materials. Subjects evaluated the performance of the auditors by assessing the sufficiency of the evidence gathered by the auditors. This measure of performance was chosen because of alleged deficiencies in evidence sufficiency in 67% of the cases against auditors in a review of SEC cases involving fraudulent financial reporting (Treadway Commission, 1987). Subjects were provided with the following statement:

> Brown and Daniels (the auditors) gathered sufficient evidence to support their opinion that the inventory of Quest at December 31, 1992 was fairly stated.

Subjects indicated how strongly they agreed with the above statement on an eleven-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree." The responses to this question will be used to test the first hypothesis.

Mediating Variables

Three mediating variables were included in the experiment: evidence recall, evidence interpretation, and evidence weighting. Measures of these variables were obtained in tasks five, six and seven, respectively.

In Task 5, subjects were provided a large t-account. On the left side they were asked to record all of the case facts they could recall that would cause them to believe the auditor should have suspected management of misstating the balance of inventory in the financial statements (FCUES). On the right side they were asked to record all of the facts they could recall that would cause them to believe that the inventory balance was fairly stated (NFCUES). The recall measure used in this experiment was the net number of fraud cues listed by the subject and was calculated as the number of items indicative of management fraud listed on the left side of the t-account less the number of items not indicative of management fraud listed on the right side of the t-account (NFCUES-FCUES).[9]

The evidence interpretation task (Task 6) provided the subjects with four items of evidence from the case that were somewhat ambiguous. Three of the items involved some form of misstatement in the physical inventory that had been uncovered by the auditor, while the fourth item related to missing documents. The four items that the subjects were asked to interpret were carefully selected to provide some ambiguity to allow for multiple interpretations. For example, the first item subjects were asked to interpret is reproduced below:

> Quest sells approximately 4,200 different items although most stores do not carry all items. At the physical inventories, Brown and Daniels double-checked the counts of some of the items. The results are summarized below:
>
> | | No. Items Double-Checked | No. Correct | No. in Error |
> |---|---|---|---|
> | Store No. 1 | 85 | 81 | 4 |
> | Store No. 2 | 100 | 92 | 8 |
> | Warehouse | 90 | 84 | 6 |
>
> All of the counts that were in error, except for one, were off by less than 5% of the correct quantity. One of the counts at the warehouse was off by 12%.

On one hand, subjects could interpret these results as providing strong evidence that inventory was fairly stated because almost all items were properly counted and those that were misstated were not misstated by significant amounts. On the other hand, subjects might interpret this item as providing little proof that the inventory was fairly stated since only two of the company's 48 stores and

the warehouse had been observed by the auditors and less than 1% of the items were double-checked at these stores. Given that such limited tests uncovered a number of minor errors and one significant error, subjects might conclude that these tests provide very little evidence that the inventory is fairly stated, or alternatively, constitute evidence that inventory was misstated.

Subjects were requested to interpret each of the items along two different dimensions. First, subjects were asked to interpret each of four items in terms of whether it provided evidence that the inventory balance was fairly stated. Each of the items was rated on an eleven-point Likert scale with endpoints ranging from "Strong Evidence that Inventory was Misstated" to "Strong Evidence that Inventory was Fairly Stated." Second, subjects were presented with the following question:

> To the extent that the last item caused you to believe inventory may have been misstated, do you believe the misstatement was probably intentional or unintentional?

Subjects responded to the above question on an eleven-point Likert scale ranging from "Intentional" to "Unintentional." Thus, subjects first interpreted whether the item was indicative of inventory misstatement, and second whether the misstatement was intentional or unintentional.[10] For each of the four items, a measure of *intentional misstatement* was calculated by multiplying the response to the first question (likelihood of misstatement) by the response to the second question (degree of intent). This measure of intentional misstatement was summed over the four items evaluated to give one overall evidence interpretation measure.

Task 7 was an evidence weighting task. Subjects were presented with four items from the case and were asked to indicate how much each of the facts affected their performance evaluation decisions. Subjects responded on an eleven-point scale ranging from "No Effect on Decision" to "Very Significant Effect on Decision." Two of these items were consistent with management fraud (merchandise in the repair area had been counted even though the merchandise belonged to customers, and the inventory in the camera department had been counted twice) while two were not indicative of management fraud (the inventory levels were very reasonable in comparison to both the prior year inventory levels and industry averages, and the manager voluntarily pointed out slow-selling merchandise to the auditors). Cues consistent with the reported outcome were expected to be weighted more heavily in hindsight than in foresight. Therefore, cues consistent with management fraud were expected to have a more significant effect on the decisions of subjects told that management fraud had occurred than subjects told that fraud had not occurred. On the other hand, just the opposite was expected for the cues that were not indicative of management

fraud. To obtain an overall weighting measure, the weights assigned to the fraud consistent cues were reverse coded and the responses were summed across the four questions.[11]

EMPIRICAL RESULTS

Table 1 provides descriptive statistics by condition, prior to standardization, for the three hypothesized mediating variables as well as the performance evaluation measure. The evidence recall measure reflects the net number of cues recalled consistent with management fraud. The evidence interpretation and evidence weighting measures reflect mean responses to the interpretation and evidence weighting items (see the previous section for additional details regarding measurement). For evidence interpretation, the theoretical range for each item is from 1 to 121 and for evidence weighting the theoretical range is from 1 to 11. The dependent variable, performance evaluation, also has a theoretical range from 1 to 11. For evidence recall, the mean of the bad-outcome group was expected to exceed the mean of the no-outcome group, and the mean of the good-outcome group was expected to be less than the mean of the no-outcome group. For evidence interpretation, evidence weighting and performance evaluation, the means of the bad-outcome group were expected to be lower than the means of the no-outcome group, while the means of the good-outcome group were expected to exceed the means of the no-outcome group.

Table 1. Descriptive Statistics for Mediating and Dependent Variables.

Condition	*N*	Evidence Recall*	Evidence Interpretation**	Evidence Weighting***	Performance Evaluation
		Mean Ratings (standard deviations in parentheses)			
Bad-Outcome Group	20	0.50 (1.82)	29.39 (14.40)	5.31 (1.35)	5.30 (2.43)
No-Outcome Group	22	1.64 (2.11)	43.88 (29.57)	5.65 (1.94)	7.23 (2.91)
Good-Outcome Group	22	0.77 (2.11)	48.36 (24.85)	6.13 (1.93)	7.45 (2.56)

* Net number of cues recalled consistent with management fraud (number of cues recalled that were consistent with management fraud less the number of cues recalled not consistent with management fraud).

** Average per-item measure of intentional misstatement; the theoretical range of each item is from 1 to 121.

*** Average measure of evidence weighting; the theoretical range of each item is from 1 to 11.

Comparison of the means for evidence interpretation, evidence weighting and performance evaluation by condition reveals that the means for each of these three variables are all in the "expected direction." In other words, for each of these variables the mean of the bad-outcome group was lower than the mean of the no-outcome group, and the mean of the good-outcome group exceeded the mean of the no-outcome group. However, this was not the case with the evidence recall measure. For evidence recall, the mean of the no-outcome group exceeds the means of both the good-outcome and bad-outcome groups.

The correlations among the variables are reported in Table 2. Several correlations among the independent, mediating and dependent variables are significant, as would be expected given the hypotheses proposed. In addition, one of the correlations among the mediating variables was significant and another was marginally significant. The correlation between evidence recall and evidence interpretation was -0.34, which was significant at the 0.01 level. The correlation between evidence recall and evidence weighting of -0.23 was marginally significant ($p = 0.07$). These negative correlations are not particularly surprising, however, since they indicate that individuals who recalled more (net) fraud consistent cues were more likely to interpret and weight the evidence in a manner consistent with management fraud.

H_1 predicts that a juror's evaluation of an auditor's performance will be affected by outcome knowledge with evaluations being adversely affected by bad-outcome information and favorably affected by good-outcome information. As noted earlier, this hypothesis is essentially a replication of prior work in a

Table 2. Correlation Matrix.

	Outcome Information	Evidence Recall	Evidence Interpretation	Evidence Weighting	Performance Evaluation
Outcome Information	1.00				
Evidence Recall	0.05 (0.71)	1.00			
Evidence Interpretation	0.31 (0.01)	-0.34 (0.01)	1.00		
Evidence Weighting	0.19 (0.14)	-0.23 (0.07)	0.15 (0.23)	1.00	
Performance Evaluation	0.31 (0.01)	-0.16 (0.20)	0.52 (0.01)	0.42 (0.01)	1.00

Note: significance levels in parentheses.

management fraud context. However, the results of this hypothesis are quite important since testing hypotheses two through seven would be illogical if the first hypothesis is not supported. Subjects' responses regarding the sufficiency of evidence gathered by the auditors were used to test the first hypothesis, and Table 1 provides descriptive statistics for the performance evaluation measure by condition. As hypothesized, the average performance evaluation was highest for those subjects receiving the good outcome and lowest for those receiving the bad outcome. The mean performance evaluation of the no-outcome group fell between the good- and bad-outcome means.

Analysis of variance (ANOVA) and planned contrasts were used to test H_1. The results of the ANOVA are reported in Table 3, which show that the main effect, outcome information, is significant at the 0.021 level. This provides strong support that outcome information does affect performance evaluation. To determine whether the means of the bad-outcome and good-outcome groups were significantly different than the no-outcome group, planned contrasts were used. The contrast of the means of the bad-outcome and the no-outcome groups was significant ($t = 2.36$, $p < 0.011$, one-tailed); however, the contrast between the no-outcome and good outcome groups was not significant ($t = 0.28$, $p < 0.39$, one-tailed). Thus, while bad-outcome information had a significant adverse effect on performance evaluations, a similar significant, favorable effect was not observed when good-outcome information was reported. Thus, H_1 is supported for bad-outcome information, but not for good-outcome information. This is consistent with much of the prior research in other contexts that has found that the effect of bad outcome information is generally greater than that of good outcome information (Hawkins & Hastie, 1990).

Path analysis was used to test the remaining hypotheses. Using path analysis, the total effect of one variable on another can be decomposed into its direct and indirect effects. The direct effect is the portion of the total effect that is not mediated or transmitted by one or more other variables, while the indirect effect is the portion of the total effect that is mediated or transmitted by other variables.[12] For purposes of the path analysis, OUTCOME was converted to an

Table 3. ANOVA Findings for Evaluation of Auditor Performance.

Source of Variation	Sum of Squares	df	Mean Squares	*F*-Value	*P*-value
Outcome	57.84	2	28.92	4.13	0.021
Error	427.52	61	7.01		
Total	485.36	63			

ordinal scale with the values of OUTCOME being 1 for subjects included in the bad-outcome condition, 2 for subjects in the no-outcome condition, and 3 for subjects in the good-outcome condition, similar to previous research (Casper et al., 1989). Additionally, since the ranges and means of the variables are quite heterogeneous due to the various ways in which the variables were measured and aggregated, all variables were standardized to aid in comparison of effect sizes between variables (Hatcher 1994).

As noted earlier in the discussion of H_1, outcome information had a significant effect on performance evaluation. Figure 2 shows the results of a very simple path analysis where performance evaluation is determined solely by outcome information. The path coefficient of 0.31 is highly significant ($t = 2.59$, $p < 0.01$, one-tailed) and indicates that a one standard deviation change in outcome information results in a change of approximately one-third of a standard deviation in performance evaluation. In such a simple model with only one dependent and one independent variable, the path coefficient is equivalent to a correlation between the two variables and the R^2 of the model reveals that 9.61% of the variance in performance evaluation can be explained by the independent variable. Clearly, there is a strong, significant relationship between these two variables.

The conceptual model is expanded in Fig. 3 to include the effects of the three mediating variables: evidence recall, evidence interpretation and evidence weighting.[13] The Goodness of Fit Index (GFI) for this model was 0.92, which is indicative of an acceptable fit of the model to the data. As shown in the figure and summarized in Table 4, evidence interpretation and evidence weighting play important mediating roles, but evidence recall does not. The path coefficient between outcome information and evidence recall is not significant ($t = 0.38$, $p = 0.35$, one-tailed). Thus, H_2, which posits that outcome information will affect the nature of the evidence recalled, is not supported. Furthermore, the path between evidence recall and performance evaluation is also insignificant ($t = 0.58$, $p = 0.28$, one-tailed). Therefore, H_3, which predicts that the nature of the evidence recalled would influence the participant's

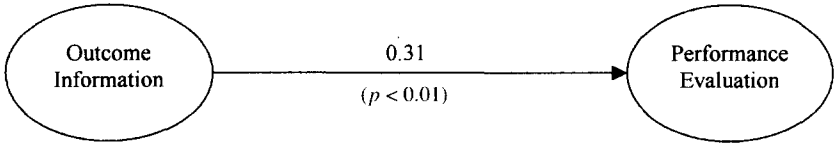


Fig. 2. Effect of Outcome Information on Performance Evaluation.

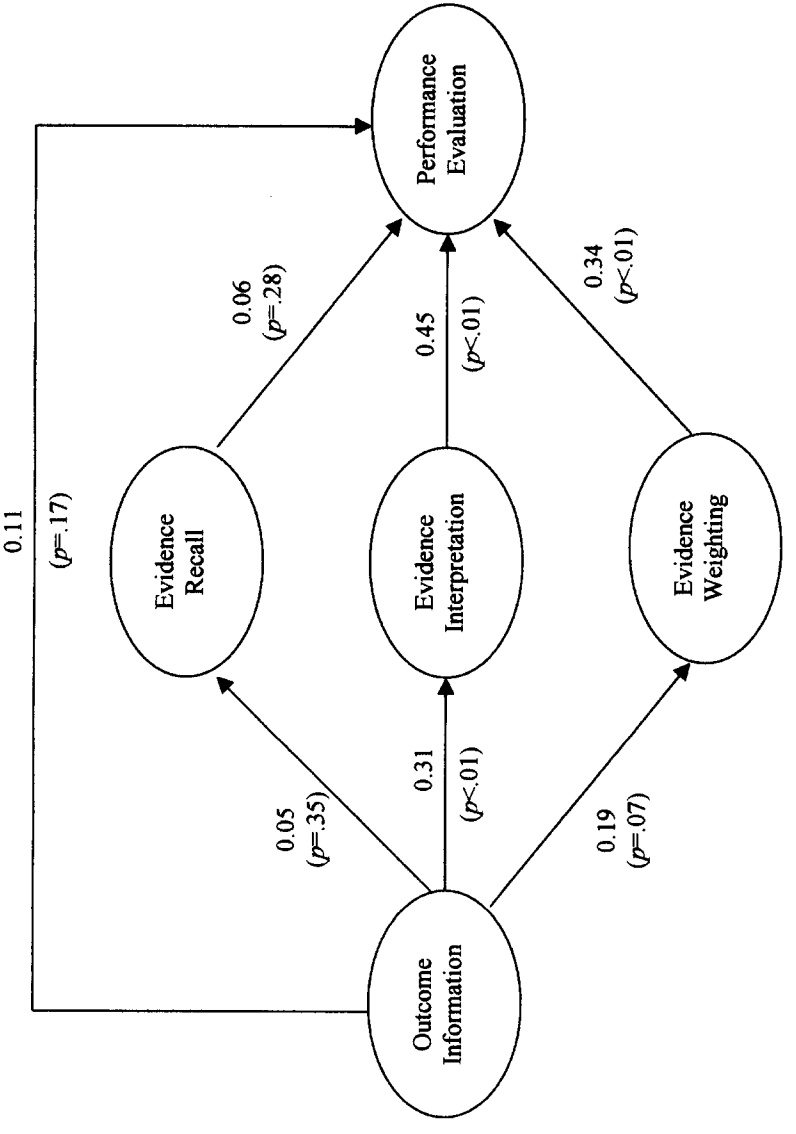


Fig. 3. Mediating Effects of Evidence Recall, Interpretation and Weighting.

Table 4. Interpretation of Effects in a Model of Auditor Performance Evaluation.

Dependent Variable	Predictor Variables	Total Effect	Indirect Effects Via			Direct Effects
			Evidence Recall	Evidence Interpretation	Evidence Weighting	
Performance Evaluation	Outcome	0.31	0.00	0.14	0.06	0.11
	Evidence Recall	0.06	—	—	—	0.06
	Evidence Interpretation	0.45	—	—	—	0.45
	Evidence Weighting	0.34	—	—	—	0.34

performance evaluation, is likewise not supported. Finally, as noted in Table 4, the indirect effect of outcome information via evidence recall, calculated as the product of the two previously mentioned path coefficients, is negligible (less than 0.01). This should not be surprising since neither H_2 nor H_3 was supported.

In contrast to evidence recall, evidence interpretation plays an important mediating role between outcome information and performance evaluation. The path coefficient between outcome information and evidence interpretation of 0.31 is significant ($t = 2.54$, $p < 0.01$, one-tailed) providing strong support for H_4. In addition, the path coefficient of 0.45 between evidence interpretation and performance evaluation is also highly significant ($t = 3.97$, $p < 0.01$, one-tailed), which provides strong support for H_5. Furthermore, as shown in Table 4, the indirect effect of outcome information via evidence interpretation, which is the product of the two path coefficients, is 0.14. Thus overall, 45.2% of the total effect of outcome information in this study is due to its indirect effect via evidence interpretation.[14] The outcome information clearly affected the interpretation of evidence, which in turn affected the evaluation of the performance of the auditor.

Evidence weighting also appears to play an important mediating role. The path coefficient between outcome information and evidence weighting is 0.19, and H_6, which hypothesizes that outcome information will affect the weighting of evidence, is supported at the 0.07 level ($t = 1.50$, one-tailed). The path coefficient of 0.34 between evidence weighting and performance evaluation was highly significant ($t = 3.23$, $p < 0.01$, one-tailed). This provides strong support for H_7, which predicts that the weighting of the evidence would affect

the participants' evaluation of the auditor. Furthermore, Table 4 shows that the indirect effect of outcome information on performance evaluation due to evidence weighting, which is the product of the two path coefficients, is 0.06. This is approximately one-fifth of the total effect of the outcome information on performance evaluation. Outcome information affected the weighting of the evidence, and the weighting of the evidence significantly affected the performance evaluation of the auditor.

Finally, the effect of including the mediating variables on the direct relationship between outcome information and performance evaluation is important to note. As can be seen from Fig. 3, after including the mediating variables the path coefficient between outcome information and performance evaluation fell to 0.11 and was no longer significant at conventional levels. Clearly, the mediators were very important in explaining the relationship between outcome information and performance evaluation.

DISCUSSION

As in prior research in accounting and psychology, this study finds that outcome information has an important effect on performance evaluations. Bad-outcome information had a significant adverse effect on the auditors' performance evaluation, while the effect of the good-outcome information was less pronounced. This does not bode well for the accounting profession as it shows that auditors are likely to be penalized for adverse outcomes, but will not benefit in a symmetrical manner from good outcomes.

The results of the path analysis reported in Fig. 3 and Table 4 help explain the outcome information effect found in prior research. Evidence interpretation and evidence weighting play an important mediating role in the relationship between outcome information and performance evaluation, while the effect of evidence recall is negligible. Subjects informed of the outcome of the case interpreted the information to be consistent with the outcome reported. In effect, this stacks the deck against the auditor when an adverse outcome occurs. Once the juror knows that management fraud occurred, the juror is likely to interpret the evidence as being indicative of management fraud. In such situations, jurors may have a difficult time understanding why the auditor could not see the red flags that are so readily apparent to them, ex post.

An analogous situation appears to apply for evidence weighting. The weighting of trial evidence appears to be related to the outcome reported. Once a juror is aware that management fraud occurred, the juror weights the evidence indicative of management fraud more heavily than evidence that is inconsistent with fraud. Again, the deck appears to be stacked against the auditor. Jurors

may be unable to understand ex post why the auditors attached so little weight, or paid so little attention to, evidence that the juror believes was clearly very important.

Overall, this study shows that in this context a large portion of the outcome information effect, approximately two-thirds in this study, is due to the indirect effects of outcome information via the interpretation and weighting of trial evidence.[15] Including the mediators in the path analysis significantly reduced the direct effect of outcome information on performance evaluation, and the path coefficient for the direct effect was no longer significant at traditional levels when the mediating variables were included. On the other hand, outcome information appeared to have very little effect on the subjects' recall of case facts. In other words, subjects recalled similar information regardless of the outcome reported. What really mattered was how they interpreted this information and how much weight they attached to specific pieces of evidence – and this was significantly affected by the reported outcome.

LIMITATIONS

As with most experiments, there are limitations to the results reported here. First, the subjects in this study worked in a controlled setting that was designed primarily to enhance the study's internal validity. The case instrument consisted of a narrative followed by various tasks where subjects responded in writing. This is different than actual trial settings where the trial evidence is presented in an oral format using witnesses. Second, the whole experiment took the average subject just over 30 minutes whereas a trial involving complex auditor liability issues could last several weeks or more. Third, the study employed student subjects from one large Midwestern university rather than subjects drawn from the general population of eligible jurors. While keeping the above limitations in mind when interpreting the results of this study is important, the results of Bornstein (1999) also seem relevant here. In a review of jury simulation research over a 20-year period, Bornstein (1999, p. 88) found that "despite the variety of approaches to conducting jury simulation research, few differences have been found as a function of either who the mock jurors are or how the mock trial is presented." Finally, the study required the measurement of multiple constructs (performance evaluation, evidence recall, evidence interpretation and evidence weighting) during the same experiment. As in any behavioral experiment measuring more than one construct, the results presented here are limited to the extent that the measurement of particular constructs may have influenced subsequent measurements of constructs.

FUTURE RESEARCH

These results have implications for three major lines of research. First, future research should examine ex ante strategies that can be employed by auditors to minimize ex post re-interpretation and re-weighting of evidence by jurors, thereby reducing potential ex post outcome effects. For example, what effect does workpaper documentation practices, workpaper review practices, or concurring opinions have on the propensity of jurors to second-guess auditors? Is it possible to document evidential matter in such a manner that results in interpretation and weighting by the juror in a manner consistent with that used by the auditor? Does the use of a well-documented, structured audit methodology reduce re-interpretation and weighting of evidence? Does the application of quantitative methods reduce re-interpretation and weighting (i.e. statistical vs. non-statistical sampling)?

Second, future research should examine the effect of formal education and task complexity on the interpretation and weighting of audit evidence. Prior research has found that expertise in a particular domain reduces, but does not eliminate, the effect of outcome information on decisions in that domain (Christensen-Szalanski & Willham, 1991). For example, doctors were less likely to show outcome bias in medical experiments than non-doctors. However, very little prior research has addressed task complexity or educational level in assessing outcome effects. This is very important because most prospective jurors will have no expertise in accounting and auditing, and those that do are likely to be excused from jury service in the voir dire process (Gobert & Jordan, 1990). Accordingly, educational level may become very important in that it may determine how well jurors will be able to understand the complex issues involved in the case. This increased capacity to understand the information may affect the degree to which subjects use outcome information to interpret and weight trial evidence. For example, Palmrose (1991) notes that of a sample of cases going to trial, the auditors' success rate was 83% on "judge trials," whereas on "jury trials" the auditors' success rate was 56%. This difference may possibly be due to the manner in which outcome effects interact with educational level via the interpretation and weighting of evidence. Judges may be more capable of judging complex cases on their merits than less educated jurors, who may be forced to resort to the use of outcome information in interpreting and weighting evidence. In addition, Anderson and Reckers (1998) found that educational level was an important factor affecting the efficacy of a particular debiasing strategy, underscoring the importance of educational level in understanding outcome effects. Further research in this area would be very useful.

Third, while several prior studies in accounting and psychology have examined the ex post mitigation of outcome effects, results have been mixed (Anderson et al., 1997; Arkes et al., 1988; Davies, 1987; Kadous, 2000; Kennedy, 1995; Lowe & Reckers, 1994). When mitigation did occur, generally it was only partial, not complete elimination. This study suggests that in the context of auditor liability, mitigation strategies that focus on evidence interpretation and weighting should be most effective. Perhaps some courtroom strategies are more useful than others in reducing the extent of re-interpretation (or re-weighting) of evidence by the jurors. While such research would be of a legal nature, results would clearly be of interest to an accounting audience.

NOTES

1. Litigation has also been linked to several other harmful, non-monetary effects on the profession. Krishnan and Krishnan (1997) provide evidence that the risk of litigation motivates auditors to withdraw from high-risk engagements, effectively making audit services more difficult to obtain for the segment of the market that most needs them. Dalton, Hill, and Ramsay (1997) found that litigation against auditors contributed to voluntary partner/manager turnover in Big 6 firms, and Hill, Metzger and Wermert (1994) argue that such turnover is likely to result in reduced audit quality. In addition, many smaller CPA firms have eliminated their audit practices due to the litigation threat making it more difficult and more costly for small businesses to obtain audits (*CPA Journal*, 1990).

2. The model discussed in this section applies to the juror's predeliberation judgment. However, prior research has shown that if a majority of jurors have the same opinion before deliberation begins, that opinion will likely carry through to the jury's verdict (Kalven & Zeisel, 1966; Pennington & Hastie, 1990).

3. For additional examples of decomposing the judgment process into subtasks, see Anderson (1981), Crocker (1981), Hastie and Park (1986) and Hogarth (1980).

4. This use of outcome groups is generally consistent with prior research in accounting (Brown & Solomon, 1987, 1993) and other fields (Hawkins & Hastie, 1990), although not all prior research has utilized a no-outcome group (Helleloid, 1988; Lipe, 1993).

5. Gobert and Jordan (1990) note that jurors were once selected based on their personal knowledge of the matters relevant to a trial; however, today such knowledge is likely to result in the juror being excused. Accordingly, non-business majors were solicited to participate so that they would have little knowledge of the matters presented in the case.

6. The experiment also included two tasks that measured the attitudes of the subjects regarding management fraud and auditors. These attitudes had no effect on the results reported in this paper. For expositional efficiency, the attitude information has not been included in the analysis of the results.

7. The quiz questions determined whether the subject had grasped the following concepts from the background information: management fraud is the intentional misstatement

of the financial statements by management, a CPA gathers evidence on a test basis during an audit, audits provide reasonable assurance, financial statements are the responsibility of management, audits are designed to detect only material errors, and that the auditors should seek evidence to corroborate information obtained through management inquiry.

8. A consumer electronics chain was used as the auditors' client since it was believed that most subjects would be able to easily visualize the types of inventory such a store would carry (e.g. televisions, VCRs, cordless telephones, etc.). The audit of inventories was chosen because this has been an audit area in which several very famous frauds have occurred (e.g. Phar-Mor, Miniscribe, Mattel). Furthermore, the Professional Issues Task Force of the AICPA Division for CPA Firms issued a Practice Alert on the physical observations of inventories (American Institute of Certified Public Accountants, 1994). Such Practice Alerts are designed to help practitioners improve the efficiency and effectiveness of their audits and are prepared based upon the experience of individual members of the task force and matters arising from litigation and peer reviews. The issuance of a Practice Alert by the AICPA Division for CPA Firms in 1994 indicates that the detection of fraud during physical inventories is a very important issue in the eyes of the profession.

9. The net number of fraud-consistent cues was used as the recall measure because it was expected that this measure would maximize the recall differences between outcome groups. Additional analyses that examined the effect of outcome on the number of fraud-consistent cues (FCUES) and fraud-inconsistent cues (NFCUES) were also performed. Outcome information did not have a significant effect on either of these variables.

10. After interpreting each of the four individual items on the two scales above, the subjects were asked to interpret the evidence collected by the auditor, considered as a whole, using the same two scales. Using the responses to these "global" interpretation questions in lieu of the individual interpretation responses has no significant effect on the results reported in this study.

11. After reverse coding the responses to the fraud-consistent cues, it was expected that the responses of subjects receiving the bad-outcome information would be lower than the responses of the subjects receiving the good-outcome information.

12. Pedhazur (1982) provides an excellent summary of path analysis. In addition, Alwin and Hauser (1975) provide an excellent discussion of the decomposition of direct and indirect effects using path analysis.

13. The path analyses in Figs 2 and 3 were estimated using a series of linear regression analyses (Pedhazur, 1982, p. 582). This solution was verified with the PROC CALIS procedure in SAS, which was also used to determine the fit of the overall model (Hatcher 1994, Chap. 4).

14. The portion of the total effect of outcome information due to its indirect effect via evidence interpretation (45.2%) is calculated by dividing the amount of the indirect effect (0.14) by the total effect (0.31).

15. The indirect effects of evidence interpretation and weighting were 0.14 and 0.06, respectively. Thus, the portion of the total effect (0.31) that was accounted for by the total of the indirect effects (0.20) was 64.5%.

16. As discussed in note 6, the experiment included two tasks that measured attitudes. These measures did not have a significant effect on the results, and accordingly, they were not incorporated into the analysis of the results. These two tasks are not shown in the abbreviated version of the instrument. In addition, for tasks six (evidence interpretation) and seven (evidence weighting), the response scales are only shown for the first question in each task.

ACKNOWLEDGMENTS

This paper is based on my Ph.D. dissertation at Indiana University. I am especially grateful to the members of my dissertation committee: Jamie Pratt (Chairman), John Hill, Edward Hirt, and Michael Metzger. In addition, the helpful comments of the editor and two anonymous reviewers are gratefully acknowledged, as are the valuable comments of D. Jordan Lowe, Robin Roberts and participants at accounting workshops at Indiana University, Louisiana State University, Iowa State University, the University of Maryland at College Park, Drake University, the American Accounting Association annual meeting, and the mid-year meeting of the AAA Auditing section.

REFERENCES

Alwin, D., & Hauser, R. (1975). The decomposition of effects in path analysis. *American Sociological Review, 40*(February), 37–47.

American Institute of Certified Public Accountants (1994). *Auditing Inventories – Physical Observations*. Practice Alert No. 94-2. New York: Division for CPA Firms – Professional Issues Task Force.

Anderson, N. (1974). Information Integration Theory: A Brief Survey. In: D. Krantz, R. Atkinson, R. Luce & P. Suppes (Eds), *Contemporary Development in Mathematical Psychology* (Vol. 2). San Francisco: Freeman.

Anderson, N. (1981). *Foundations of Information Integration Theory*. New York: Academic Press.

Anderson, J., Jennings, M., Lowe, D. J., & Reckers, P. (1997). The mitigation of hindsight bias in judges' evaluation of auditor decisions. *Auditing: A Journal of Practice and Theory, 16*(2), 20–39.

Anderson, J., & Reckers, P. (1998). Mitigating hindsight bias in jurors' evaluation of auditor decisions: Considering alternative outcomes and the education of jurors. *Advances in Accounting, 16*, 221–237.

Arkes, H., Faust, D., Guilmette, T., & Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology, 73*(2), 305–307.

Baron, J., & Hershey, J. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology, 54*, 569–579.

Berton, L. (1999a). Firms still coughing up the dough. *Accounting Today*. February 8, p. NA.

Berton, L. (1999b). Condemned to repeat accounting mistakes past. *Accounting Today*, October 25, p. NA.

Bornstein, B. H. (1999). The ecological validity of jury simulations: Is the jury still out? *Law and Human Behavior, 23*(1), 75–91.

Brown, C., & Solomon, I. (1987). Effects of outcome information on evaluations of managerial decisions. *The Accounting Review, LXII*(3), 564–577.

Brown, C., & Solomon, I. (1993). An experimental investigation of explanations for outcome effects on appraisals of capital-budgeting decisions. *Contemporary Accounting Research, 10*(1), 83–111.

Buchman, T. (1985). An effect of hindsight on predicting bankruptcy with accounting information. *Accounting, Organizations and Society, 10*(3), 267–285.

Casper, J., Benedict, K., & Kelly, J. (1988). Cognition, attitudes and decision-making in search and seizure cases. *Journal of Applied Social Psychology, 18*, 93–113.

Casper, J., Benedict, K., & Perry, J. (1989). Juror decision making, attitudes, and the hindsight bias. *Law and Human Behavior, 13*, 291–310.

Christensen-Szalanski, J., & Willham, C. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes, 48*, 147–168.

Cloyd, C. B., Frederickson, J. R., & Hill, J. W. (1998). Independent auditor litigation: Recent events and related research. *Journal of Accounting and Public Policy, 17*, 121–142.

CPA Journal (1990). Lawsuit fears forcing auditors to cut services. *LX*(December), 6.

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin, 90*, 272–292.

Dalton, D. R., Hill, J. W., & Ramsay, R. J. (1997). The threat of litigation and voluntary partner/manager turnover in Big Six firms. *Journal of Accounting and Public Policy, 16*, 379–413.

Davies, M. F. (1987). Reduction of hindsight bias by restoration of foresight perspective: Effectiveness of foresight-encoding and hindsight retrieval strategies. *Organizational Behavior and Human Decision Processes, 40*, 50–68.

Dellarosa, D., & Bourne, L., Jr. (1984). Decisions and memory: Differential retrievability of consistent and contradictory evidence. *Journal of Verbal Learning and Behavior, 23*, 669–682.

Fischhoff, B. (1975). Hindsight . . . foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology, 1*, 288–299.

General Accounting Office (1989). *CPA Audit Quality: Failures of CPA Audits to Identify and Report Significant Savings and Loan Problems.* Washington: General Accounting Office. GAO/AFMD-89-45.

Gobert, J., & Jordan, W. (1990). *Jury Selection: The Law, Art, and Science of Selecting a Jury.* (2nd ed.). New York: McGraw-Hill.

Hastie, R. (1993). Algebraic models of juror decision-making processes. In: R. Hastie (Ed.), *Inside the Juror.* New York: Cambridge University Press.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is online or memory-based. *Psychological Review, 93*, 258–268.

Hatcher, L. (1994) *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling.* Cary, NC: SAS Institute.

Hawkins, S., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin, 107*, 311–327.

Hearing Before the Committee on Banking, Finance and Urban Affairs, House of Representatives (1989). *Failure of Independent CPA's to Identify Fraud, Waste and Mismanagement and Assure Accurate Financial Position of Troubled S&L's.* Washington: U.S. Government Printing Office.

Hearings Before the Subcommittee on Reports, Accounting and Management of the Committee on Governmental Affairs, U.S. Senate (1977). *Accounting and Auditing Practices and Procedures.* Washington: U.S. Government Printing Office.

Helleloid, R. (1988). Hindsight judgments about taxpayers' expectations. *The Journal of the American Taxation Association, 9*(2), 31–46.

Hershey, J., & Baron, J. (1992). Judgment by outcomes: When is it justified? *Organizational Behavior and Human Decision Processes, 53*, 89–93.

Hershey, J., & Baron, J. (1995). Judgment by outcomes: Why it is interesting? A reply to Hershey and Baron: Judgment by outcomes: When is it justified? *Organizational Behavior and Human Decision Processes, 62*(1), 127.

Hill, J. W., Metzger, M. B., & Wermert, J. G. (1994). The spectre of disproportionate auditor liability in the savings and loan crisis. *Critical Perspectives on Accounting*, *5*(2), 133–177.

Hogarth, R. (1980). *Judgment and Choice.* New York: Wiley.

Kadous, K. (2000). The effects of audit quality and consequence severity on juror evaluations of auditor responsibility for plaintiff losses. *The Accounting Review*, *75*(3), 327–341.

Kalven, H., Jr., & Zeisel, H. (1966). *The American Jury.* Boston: Little, Brown.

Kaplan, M., & Miller, L. (1978). Reducing the Effects of Juror Bias. *Journal of Personality and Social Psychology*, *36*(12), 1443–1455.

Kennedy, J. (1995). Debiasing the curse of knowledge in audit judgment. *The Accounting Review*, *70*(2), 249–273.

Kinney, W., Jr. (1993). Auditor's liability: Opportunities for research. *Journal of Economics and Management Strategy*, *2*(3), 349–360.

Kinney, W., Jr. (1994). Audit litigation research: Professional help is needed. *Accounting Horizons*, *8*(2), 80–86.

Krishnan, J., & Krishnan, J. (1997). Litigation risk and auditor resignations. *The Accounting Review*, *72*(4), 539–560.

Lipe, M. (1993). Analyzing the variance investigation decision: The Effects of Outcomes, Mental Accounting, and Framing. *The Accounting Review*, *68*(4), 748–764.

Lowe, D. J., & Reckers, P. (1994). The effect of hindsight bias on jurors' evaluations of auditor decisions. *Decision Sciences*, *25*(3), 401–426.

MacDonald, E. (1999). SEC to boost accounting-fraud attack, Work more with criminal prosecutors. *The Wall Street Journal*, (December 8), A6.

Mednick, R. (1996). Is the auditor liable as business agent or a professional? Genesis and resolution of the liability crisis from a primarily U.S. perspective. Speech delivered at The Sixth Jerusalem Conference on Accountancy. November 12.

Ostrum, T., Werner, C., & Saks, M. (1978). An integration theory analysis of jurors' presumptions of guilt or innocence. *Journal of Personality and Social Psychology*, *36*(4), 436–450.

Palmrose, Z. (1991). Trials of legal disputes involving independent auditors: Some empirical evidence. *Journal of Accounting Research*, *29*(Supplement), 149–185.

Palmrose, Z. (1997a). Audit litigation research: Do the merits matter? An Assessment and Directions for Future Research. *Journal of Accounting and Public Policy*, *16*, 355–378.

Palmrose, Z. (1997b). Who got sued? *Journal of Accountancy*, (March), 67–69.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction.* (2nd ed.). New York: Holt, Rinehart & Winston.

Pennington, N., & Hastie, R. (1981). Juror decision-making models: The generalization gap. *Psychological Bulletin*, *89*, 246–287.

Pennington, N., & Hastie, R. (1990). Practical implications of psychological research on jury decision making. *Personality and Social Psychology Bulletin*, *16*, 90–105.

Prosser and Keeton on Torts (1977). (5th ed.), Section 170. West publishing Co. St Paul, MN.

Sand, L., Siffert, J., Loughlin, W., Reiss, W., & Batterman, N. (1997). *Modern Federal Jury Instructions* (Vol. 4). San Francisco: Matthew Bender & Co.

Schroeder, M. (2001). SEC list of accounting-fraud probes grows, stretching agencies resources. *The Wall Street Journal* [On-line]. Available: http://interactive.wsj.com/articles/SB994366683510250066.htm (July 6, 2001).

Silicano, J. (1997). Trends in independent auditor liability: The Emergence of a Sane Consensus. *Journal of Accounting and Public Policy*, *16*, 339–353.

Slovic, P., & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology*, *3*, 544–551.

Staloch v. Holm (1907). 111 N. W., pp. 264–269.
Teig v. St. John's Hospital (1963). 387 P. 2nd., pp. 527–536.
Treadway Commission (National Commission on Fraudulent Financial Reporting) (1987). *Report of the National Commission on Fraudulent Financial Reporting*. AICPA: New York.

APPENDIX A

Abbreviated Version of Experimental Instrument[16]

TASK 3

The Audit of Quest Corporation

You are about to read a case that describes a portion of the 1992 audit of Quest Corporation. The case deals with the auditor's responsibility to detect management fraud (i.e. intentional misstatement of the financial statements by management). You will be given **all** of the relevant information that was available to the auditors of Quest at the time of the audit. Based upon this information, you will be asked for your opinion about the adequacy of the auditor's work to detect management fraud. Your responses will help auditors understand how individuals such as yourself evaluate their performance.

Quest's financial statements have been audited by the international auditing firm of Brown and Daniels since 1984. Brown and Daniels employs approximately 12,000 people in the United States and has offices in most major cities, including San Francisco and Los Angeles.

[*An outcome paragraph was included here that was dependent upon the treatment group of the subject. The three outcome paragraphs are provided at the end of this appendix.*]

Quest Corporation

Quest Corporation is a retailer of consumer electronics, offering approximately 4,200 different products of nationally known brands. Some of the "high ticket" items that the company sells are televisions, stereos, VCRs, camcorders, compact disc players, and personal computers. The company's 48 stores, located primarily in San Francisco and Los Angeles, had sales of $552 million for the year ended December 31, 1992.

The Physical Count of Inventory by Quest

Quest counted all of the inventory at each of its stores and its warehouse on December 31st. (Reminder: Inventory is merchandise the company has purchased

that it intends to resell at a higher price.) A count of the inventory of a company is usually referred to as a physical inventory. It is important that a physical inventory be done correctly because errors affect the amount of inventory reported in the company's financial statements as well as the earnings of the company. Quest's Accounting Department provided each store manager with specific instructions on the procedures to be followed when counting the inventory.

All stores were closed during the physical inventory to help ensure that an accurate count would be taken. All merchandise was counted by people who work at the store, or in the case of the warehouse, by people who actually work in the warehouse. Thus, the people who counted the inventory were familiar with it. As the workers counted the inventory, they prepared pre-numbered tags for each inventory item and attached the tag to the inventory item. The tags included the description, quantity, and unit of measure (units, pairs, dozens, etc). One part of this tag was collected later and used for calculating the total amount of inventory.

Audit Planning

Brown and Daniels planned the audit of Quest in July 1992. They considered the factors that affect the risk of errors in the financial statements and estimated the overall risk to be moderate (or average). Based upon this estimate, Brown and Daniels was required by professional auditing standards to collect a "moderate" (or average) amount of evidence to provide reasonable assurance.

The 1992 Audit of Inventory

Brown and Daniels spent 130 hours auditing the $71.5 million of inventory that Quest reported in its 1992 financial statements. The work done by Brown and Daniels to audit inventory is discussed below.

In prior years, Brown and Daniels found the Quest staff did a very good job of taking the physical inventory. Most of the managers of Quest stores have several years of experience and know the physical inventory procedures very well. Accordingly, Brown and Daniels decided it would not be necessary for them to attend the physical inventory at all 48 stores. Instead, the auditors decided to attend the physical inventories at two stores and the warehouse. However, Quest did not know in advance which of the stores the auditors would attend. Listed below are the results of the procedures performed by Brown and Daniels while observing the physical inventories:

- Quest sells approximately 4,200 different items although most stores do not carry all items. At the physical inventories, Brown and Daniels double-checked

the counts of some of the items. The results are summarized below:

	Number of Items Double-Checked	Number Correct	Number in Error
Store No. 1	85	81	4
Store No. 2	100	92	8
Warehouse	90	84	6

All of the counts that were in error, except for one, were off by less than 5% of the correct quantity. One of the counts at the warehouse was off by 12%.

- One of the store managers pointed out to the auditors approximately $10,000 of slow-selling merchandise from several different departments. This merchandise was assigned a very low value since it would have to be sold at very low "clearance" prices. It would have been very unlikely that the auditors would have detected this slow-selling inventory without the store manager pointing it out.
- The inventory in the camera department at one of the stores had been counted twice (and included in inventory twice). When brought to the manager's attention, he informed the auditors that this was an accident. He explained that each employee was told which areas he or she was supposed to count. The camera department was accidentally assigned to two different individuals and each counted the goods without realizing that someone else had also counted this merchandise.
- The auditors discovered that Quest had counted the inventory in the repair area (at the warehouse) even though this merchandise belonged to customers and was simply returned to Quest to be serviced. The warehouse manager told the auditors that the employee who counted this area was fairly new and did not understand the inventory instructions. The manager personally destroyed these inventory tags.

In addition to observing the physical inventory, Brown and Daniels performed the following additional audit procedures:

- Quest has an Internal Audit Department that serves as a "watchdog" over the accounting department and management. Quest's Internal Audit Department attended the inventory at five stores, all different than the stores attended by Brown and Daniels. Max Anderson, the head of this department, told Brown and Daniels that no significant errors were found at four of the five stores. However, errors totaling approximately $110,000 were found at one store due

to mistakes in recording the unit-of-measure (i.e. cases, units, dozens, etc.) For example, in one instance, 14 cordless telephones were recorded as 14 dozen cordless telephones. Anderson said he believed the error was due to inexperience on the part of the staff at this store because this stored opened during 1992.

- Brown and Daniels double-checked 30 transfers of inventory (between the warehouse and the stores) just before and after the physical inventory. Of the 30 transfers that were checked, 28 were accounted for properly. Two transfers for approximately $82,000 of inventory were double-counted. This error resulted in this inventory being included in both the store's inventory and in the warehouse's inventory. Quest management investigated the errors and determined that they were accidental. Management corrected the errors. Brown and Daniels double-checked an additional ten transfers and all were accounted for properly. The auditors decided no further investigation was necessary.
- Brown and Daniels compared the amount of inventory on hand in December 1992 with the amount of inventory owned in prior years. The amount owned in the current year increased 1.2% over the prior year. Also, Brown and Daniels compared Quest's inventories to industry averages (found in trade magazines published by the Retail Industry). The average inventory of a Quest store was 2% less than the inventory included in an average electronics store. Brown and Daniels thought Quest's inventory level appeared reasonable.
- Brown and Daniels performed tests of the dollar values Quest assigned to the inventory. Quest was not able to find the invoices (a type of document) to support the value for six of the 70 items initially picked for audit testing. Quest management said that due to the high volume of transactions processed during a month, it is not uncommon for some documents to become misplaced. Brown and Daniels selected six different items. Since these transactions all appeared proper, the auditors decided that no further investigation was necessary.
- Procedures were performed to ensure that the information from the physical inventories had been correctly summarized, that any necessary adjustments to the accounting records resulting from the physical inventories had been recorded, and that the inventory transactions close to December 31st had been recorded in the proper accounting period. Baker and Daniels found no errors when performing these procedures.

Summary

Based upon the audit procedures performed to audit inventory, Brown and Daniels concluded that the inventory amount of $71.5 million included in the 1992 Quest financial statements was fairly stated.

TASK 4

After reading the information about the 1992 audit of Quest, please indicate how strongly you agree or disagree with the following statement. Do not refer back to the material you read in Task 5.

Brown and Daniels gathered sufficient evidence to support their opinion that the inventory of Quest at December 31, 1992 was fairly stated.

Strongly Disagree										Strongly Agree
1	2	3	4	5	6	7	8	9	10	11

TASK 5

Do not refer back to the facts presented in the earlier case. Based only upon your recall of the Quest Case information, indicate those facts that

(a) cause you to believe the auditor should have suspected management of misstating the balance of inventory in the 1992 financial statements (i.e. those facts indicating that management fraud may have occurred).
(b) cause you to believe that the auditor should not have suspected management of misstating the balance of inventory in the 1992 financial statements (i.e. facts that lead you to believe that inventory is fairly stated):

You are not required to list an equal number of items under each column.

(a) Facts indicating Management Fraud	(b) Facts Indicating Inventory Was Fairly Stated

TASK 6

Please indicate how you would interpret each of the following items in terms of whether it provides evidence that the inventory balance was fairly stated.

(1) Quest sells approximately 4,200 different items although most stores do not carry all items. At the physical inventories, Brown and Daniels double-checked the counts of some of the items. The results are summarized below:

	Number of Items Double-Checked	Number Correct	Number in Error
Store No. 1	85	81	4
Store No. 2	100	92	8
Warehouse	90	84	6

All of the counts that were in error, except for one, were off by less than 5% of the correct quantity. One of the counts at the warehouse was off by 12%.

Strong Evidence that Inventory Was Misstated										Strong Evidence that Inventory was Fairly Stated
1	2	3	4	5	6	7	8	9	10	11

To the extent that the last item caused you to believe inventory may have been misstated, do you believe the misstatement was probably intentional or unintentional?

Intentional										Unintentional
1	2	3	4	5	6	7	8	9	10	11

(2) Brown and Daniels performed tests of the values Quest assigned to the inventory. Quest was not able to find the invoices (a type of document) to support the value for six of the 70 items initially selected for audit testing.

(3) Brown and Daniels double-checked 30 transfers of inventory (between the warehouse and the stores) just before and after the physical inventory. Of the 30 transfers that were checked, 28 were accounted for properly. Two transfers for approximately $82,000 of inventory were double-counted. This error resulted in this inventory being included in the store's inventory and

in the warehouse's inventory. Quest management investigated the errors and determined that they were accidental. Management corrected the errors. Brown and Daniels double-checked ten additional transfers and all were accounted for properly.

(4) Quest's Internal Audit Department found errors of approximately $110,000 at one store due to mistakes in recording the unit-of-measure (i.e. cases, units, dozens, etc.) For example, in one instance, 14 cordless telephones were recorded as 14 dozen cordless telephones. Anderson said he believed the error was due to inexperience on the part of the staff at this store because this stored opened during 1992.

TASK 7

Earlier you were asked to judge the sufficiency of evidence collected by Brown and Daniels. **You may review your answer to that question (in Task 4) at this time.** Please indicate on the scale below how much each of the following affected that decision:

(1) The auditors discovered that Quest had counted the inventory in the repair shop area (at the warehouse) even though this merchandise belonged to customers and was simply returned to Quest to be serviced.

No Effect on Decision — Very Significant Effect on Decision

1 2 3 4 5 6 7 8 9 10 11

(2) Brown and Daniels compared the amount of inventory on hand in December 1992 with the amount of inventory owned in prior years. The amount owned in the current year increased 1.2% over the prior year. Brown and Daniels also compared Quest's inventories to industry averages. The average inventory of a Quest store was 2% less than the inventory included in an average electronics store. Brown and Daniels thought Quest's inventory level appeared reasonable.

(3) The inventory in the camera department had been counted twice.

(4) The store manager pointed out to the auditors approximately $10,000 of slow-selling merchandise from several different departments. This merchandise was assigned a very low value since it would have to be sold at very

low "clearance" prices. It would have been very unlikely that the auditors would have detected this slow selling inventory without the store manager pointing it out.

Outcome Information Paragraphs

Outcome Paragraph for No-Outcome Condition

The company used for this case, Quest Corporation, was selected **randomly** from the list of clients of Brown and Daniels. It is not known whether management fraud did or did not occur at Quest during 1992. You should evaluate Brown and Daniels' performance using the information that was available to them at the time they performed the audit.

Outcome Paragraph for Bad-Outcome Condition

Quest's auditors did not detect any management fraud during the 1992 audit. However, after the 1992 audit was completed, allegations arose that Quest's management intentionally overstated the value of Quest's inventories in its 1992 financial statements. **A special fraud investigation concluded that management fraud had occurred. However, the fact that the auditors did not detect the management fraud does not necessarily imply that their performance was inadequate.** You should evaluate the auditors' performance using the information that was available to the auditors at the time they performed the audit. Please do not allow your decisions in this case to be affected by the conclusions of the special fraud investigation.

Outcome Paragraph for Good-Outcome Condition

Quest's auditors did not detect any management fraud during the 1992 audit. However, after the 1992 audit was completed, allegations arose that Quest's management overstated the value of Quest's inventories in its 1992 financial statements. **A special fraud investigation concluded that no management fraud had occurred. However, the fact that no management fraud occurred does not necessarily imply that the auditor's performance was adequate.** You should evaluate the auditors' performance using the information that was available to the auditors at the time they performed the audit. Please do not allow your decisions in this case to be affected by the conclusions of the special fraud investigation.

THE EFFECTS OF COGNITIVE LOAD ON DECISION AID USERS

Jacob M. Rose

ABSTRACT

This study employs multiple measures of schema acquisition and analyzes subjects' problem-solving error patterns in order to investigate schema acquisition by decision aid users. Results of both the error analysis and multiple schema acquisition measures indicate that: (1) problems of ordered complexity can effectively capture differences in schema acquisition of decision aid users; and (2) problem solvers rely on incorrect simple schemata to solve problems when they have not acquired the complex schemata necessary to solve a problem. The results also provide additional support for prior findings that problem-solving schemata are acquired in a linear order flowing from computationally simple problems to more complex problems, and that cognitive load interferes with the acquisition of schemata from decision aids.

INTRODUCTION

Auditors and tax professionals regularly employ decision aids on the job, and relevant experience is acquired during decision aid use. As professionals gain experience, they are able to increasingly rely on knowledge and judgment to make complex decisions. Assurance and tax firms spend significant resources on the development of decision aids to assist their professionals, and firms

Advances in Accounting Behavioral Research, Volume 5, pages 115–140.

ISBN: 0-7623-0953-9

desire aids that maximize both judgment performance and knowledge acquisition. Prior research has demonstrated that system designers and accounting firms need to consider the cognitive load imposed by decision aid design alternatives. Cognitive load is the consumption of working memory during problem solution and schema acquisition (Sweller, 1988). The design of decision aids can create cognitive load and significantly affect a user's ability to learn from the aids and develop problem-solving schemata (Rose & Wolfe, 2000). Schemata are knowledge structures and are commonly defined as "cognitive constructs that permit problem-solvers to recognize a problem as belonging to a specific category requiring particular moves for completion" (Tarmizi & Sweller, 1988). Schemata are considered to be the foundation of expertise (Sweller, 1993). A critical issue in studies of the effects of system design or instructional design on schema acquisition is the measurement of schema acquisition. The only prior study to investigate schema acquisition by decision aid users employed analogical problems with varying levels of complexity to measure schema acquisition (Rose & Wolfe, 2000). Subjects using decision aids that produced lower levels of cognitive load correctly completed more and also more complex analogical problems than subjects using aids with high levels of cognitive load. This provided evidence that subjects facing low levels of cognitive load not only acquired greater quantities of declarative knowledge but also more complex schemata.

The current research has three primary purposes. First, this study employs multiple measures of learning performance to measure schema acquisition in aided learning environments. Given that only one study has measured schema acquisition in aided environments and that problems of ordered complexity were the sole measure of schema acquisition, triangulated results using alternative measures of schema acquisition are needed. The cognitive load produced by a decision aid appears to hinder learning from the aid, but the only available evidence for the effects of cognitive load relies on manipulations of problem complexity. The suitability of using problem complexity to capture schema acquisition must be demonstrated before substantial resources are devoted to designing systems that minimize cognitive load. The use of multiple measures of schema acquisition also provides additional evidence for the detrimental effects of cognitive load when learning from decision aids, which is the second purpose of this research.

The third purpose of this study is to investigate the effects of cognitive load on aid users' development of problem solving strategies. This study employs a learning task where subjects use decision aids that produce differential levels of cognitive load to solve tax problems and learn through experience. Subjects' problem-solving error patterns are analyzed to determine the problem-solving

strategies developed by aid users facing high and low levels of cognitive load. Aid users facing high levels of cognitive load are expected to inappropriately rely on simple problem solving strategies for problems requiring more complex solution methods. In addition, the error analysis is employed to validate the appropriateness of manipulating problem complexity to capture schema acquisition by decision aid users.

Results of both the error analysis and the alternative schema acquisition measure indicate that: (1) problems of ordered complexity and text editing effectively capture differences in schema acquisition by aid users; and (2) problem solvers rely on incorrect, simple schemata to solve problems when they have not acquired the schemata necessary to solve the problems. Results also reinforce findings from prior research. Problem-solving schemata are acquired in a linear order flowing from computationally simple problems to more complex problems, and the cognitive load produced by a decision aid hinders the acquisition of schemata.

The remainder of the paper describes the relevant literature and hypothesis development, followed by a description of the methods and results. The final section includes discussion and conclusions, as well as limitations.

BACKGROUND AND HYPOTHESIS DEVELOPMENT

Human Memory, Schemata, and Problem Solving

This research is based on the following characteristics of the human cognitive system: (1) working memory is very limited (Miller, 1956; Baddeley, 1992); (2) long-term memory housing schemata is effectively unlimited (Simon & Gillmartin, 1973); (3) schema acquisition is a primary learning mechanism and source of intellectual skill (Sweller, 1993); and (4) sufficient working memory must be available to acquire schemata (Sweller, 1988, 1993; Sweller et al., 1990; Mousavi et al., 1995).

Working memory is used to hold and process current information (Baddeley, 1992). Seminal work by Miller (1956) indicates working memory's severe limits, and other research has validated this finding (Simon, 1974; Penney, 1989). Long-term memory, on the other hand, is boundless. Simon and Gillmartin (1973) estimate that in rich domains experts have literally thousands of "chunks" of information in long-term memory. This store of information seemingly defines skilled expert performance (Sweller, 1993).

Schema theory is important because it explains how information in long-term memory is organized and stored. While a number of definitions for schemata exist, Anderson (1985) defines them simply as organized knowledge of the

world, and Sweller (1993) defines them as constructs that organize knowledge in the manner in which it will be used. Compelling evidence of schema usage in higher level intellectual activities include Chase and Simon's (1973) discovery that expert chess players have and use schematic knowledge of realistic board configurations, but they do not have such schemata for improbable or impossible board configurations. Similar studies in physics find that experts have problem-solving schemata that allow them to categorize problems according to their solution mode, while novices categorize such problems by surface features (Chi, Glaser & Rees, 1982).

The significance of schemata in problem solving is often illustrated in terms of math problems (Mayer, 1981; Low & Over, 1992; Sweller, 1993). For example, in the simple algebraic equation, $ax + bx = c$, most readers will immediately begin isolating x in the left hand side of the equation: $x(a + b)$. As Sweller (1993) points out in a similar example, this is unnecessary, and no rules of algebra dictate such a first move. Schematic knowledge of how such a problem should be solved drives this "correct" solution approach. In a similar fashion, expert accountants should develop very detailed schemata for recognizing and solving accounting problems.

Weber (1980) validates this in a free recall experiment requiring auditors to recall information technology (IT) controls. Experienced IT auditors demonstrated more cue clustering than students, indicating that experienced auditors had organized the IT controls into schemata. Free recall from students was more random, apparently because they lacked relevant schemata. Libby (1985) found that potential errors identified by experienced auditors during analytical review were more common in practice than errors identified by novices. Experts had apparently developed financial statement error schemata that increased their ability to recognize reasonable errors. The literature indicates that expert accountants possess schemata that define their expertise.

Schemata enable experts to recognize problem states and relevant problem features before problem solving ever begins. In demonstrating this, Mayer (1981) classified approximately 1100 algebra word problems into specific categories; and within categories, he identified specific solution templates. Following Mayers' model, Low and Over (1990) assessed schematic knowledge based on students' abilities to identify solution templates within problem categories. From the errors they discovered, Low and Over (1990) suggested that simple templates were acquired before more complex templates, and errors were often defined as the use of a simple template in place of the correct, more complex template. In a direct test of the acquisition order of problem-solving schemata, Low and Over (1992) found evidence that simple schemata need to be in place before more complex schemata can be acquired. Their work suggests that understanding

the nature of schema acquisition is important for the ordering of instruction that attempts to create problem-solving schemata.

Cognitive Load Theory

Mayer (1979) proposed that acquiring new knowledge requires effort and the integration of new information with existing schemata or the creation of new schemata. Mayer's theory, however, did not help to explain the fact that not all experience is suitable for schema acquisition. Sweller (1988) outlined a theory of cognitive load to explain the failure of some experience to produce problem-solving schemata. The theory indicated that complex problem solving required capturing problem detail and goal states in working memory, thereby, eliminating memory space available for schema acquisition. To test this theory, research has focused on problem requirement and presentation techniques that might be expected to lower the demands on working memory, therefore enhancing schema acquisition by reducing cognitive load (Sweller & Levine, 1982; Sweller et al., 1983; Owen & Sweller, 1985; Tarmizi & Sweller, 1988).

Cognitive load theory indicates that problems create cognitive load through their structure or format when they embody a split attention effect. Attention is split when the act of solving a problem requires that information from one source be held in working memory while information from another source is evaluated (Mousavi et al., 1995). Splitting attention reduces the memory available for acquiring schemata because it requires information to be held in working memory while attention is given to other information. Any instructional materials that embody a split attention effect create cognitive load (Sweller et al., 1990; Chandler & Sweller, 1991; Chandler & Sweller, 1992). Rose and Wolfe (2000) found that the placement of explanations in a decision aid can create split attention effects and increase cognitive load. Further, they demonstrated that the cognitive load induced by explanation placement results in decreased levels of schema acquisition for decision aid users.

Measuring the Acquisition of Schematic Knowledge

Low and Over (1990) assessed schematic knowledge of secondary school students by requiring them to determine if algebraic word problems contained sufficient, missing, or irrelevant information. Upon evaluating the errors that the students made, Low and Over suggested that the students acquired problem solving templates in a linear order of complexity for the algebraic word problems. That is, students did not acquire more complex problem solving templates before they had acquired simple templates.

In another test of the acquisition order of schematic knowledge, Low and Over (1992) used a series of geometry problems that required progressively more complex problem solving templates. Their four problems required students to determine the area of a rectangle based upon: (1) information about adjacent rectangle sides; (2) information about one side and its ratio to the other; (3) information about one side and the perimeter; and (4) information about one side and the diagonal.[1] The first problem represents required knowledge for all others. That is, students must know that the multiplication of the lengths of two adjacent sides equals a rectangle's area. However, the remaining problems are independent of one another. For example, a student's knowledge of the Pythagorean theorem allows the solution of problem four and is completely disconnected from their knowledge of ratios needed in problem two. Low and Over found that student subjects failed to acquire more complex problem-solving templates if they had not acquired simpler templates first, even though knowledge of simpler templates was not needed to learn higher-level templates. Their results indicate a definite order from simple to complex in the acquisition of problem-type schemata.

Rose and Wolfe (2000) tested the acquisition order of problem-type schemata when training involves the use of an automated decision aid. They found that aid users acquired problem-solving schemata for tax problems in a simple to complex order, and increased cognitive load resulted in the acquisition of less complex schemata. Similar to Low and Over (1990), Rose and Wolfe employed a series of tax problems that required progressively more complex problem-solving templates to measure schema acquisition. To ascertain the relative complexity of the problem-solving templates in their tax decision aid, Rose and Wolfe used the number of rules required in a particular template (i.e. simpler templates had fewer rules).

To verify that the level of cognitive load produced by a decision aid hinders schema acquisition, the current study employs two measures of schema acquisition: problems that require progressively more complex problem-solving templates and text-editing problems. Solution of text-editing problems requires subjects to identify problems as belonging to a specific class and to understand the operations required to solve problems of that class. These requirements mirror the elements of a schema for problem solving.

Extending Low and Over's (1992) finding to decision aid use would suggest that subjects will not develop schemata for more intricate problems before they have acquired schemata for more basic problems through experience with an aid. Automated decision aids complete simple calculations and operations for the user and, therefore, allow users the opportunity to devote cognitive resources to more complex problem-solving procedures. Indeed, Glover et al. (1997) found

that decision aid users reduced cognitive effort even when given an aid that did not contain sufficient procedures to reach correct solutions in all cases. Decision aid users appear willing to rely on the aid to perform simple tasks, even when the aid itself is not highly reliable. Decision aid users also have the capability to attempt indirect solution paths not available in manual tasks, because automated aids can complete many steps without user intervention. As a result, learning in decision-aided environments is unlike learning in unaided environments, and aided learners may possibly develop schemata for complex problems before acquiring schemata for simple problems.

The importance of understanding the complexity and order of schema acquisition attained from decision aid use is that: (1) in instances where only simple schemata are acquired through aid use, acquiring complex schemata when completing tasks without the aid would be hindered; and (2) aids that fail to provide users with the opportunity to acquire simple schemata will reduce the ability of users to acquire more complex schemata. The issues of order and complexity of knowledge acquisition are highly relevant and require additional study because decision aids may not lead to the learning effects previously demonstrated in traditional learning environments. Further, prior research has provided only limited evidence that decision aid users acquire schemata in an order of complexity or that the cognitive load produced by decision aids interferes with schema acquisition.

The following hypotheses, stated in alternative form, are used to test the propositions that schemata are acquired in the order of complexity and that cognitive load interferes with schema acquisition. Rose and Wolfe (2000) addressed these hypotheses in prior work. The current study re-evaluates these issues using multiple measures of schema acquisition. Multiple measures will provide more convincing evidence of the effects of cognitive load on schema acquisition and the order of acquisition by aid users than is provided in the available literature.

Hypothesis 1: Users of decision aids that produce higher levels of cognitive load will acquire less complex problem-solving schemata than users of decision aids that produce lower levels of cognitive load.

Hypothesis 2: The acquisition of schematic knowledge will follow a linear order flowing from computationally simple problem solving templates to those that are more complex.

This study also analyzes the errors made by subjects during problem solving. Problem solvers who lack expertise and well-developed schemata tend to rely on heuristics and simple solution strategies when faced with complex problems (Gick, 1986; Sweller & Levine, 1982). Low and Over (1992) demonstrated that

students who had not acquired more complex algebra templates consistently made the mistake of using an inappropriate, simple template when the problem called for a more complex template. Aid users will expectedly apply simple problem-solving schemata that have been acquired to problems that require more complex schemata that have not been acquired. Revelation of the nature of the errors made during problem solving will indicate the problem-solving strategies adopted by subjects who learned through experience with a decision aid. Further, understanding the nature of errors will make predicting the errors that problem solvers make when no aid is available possible. Evidence of reliance on simple templates when more complex templates are necessary for problem solution will also provide additional evidence that problem-solving schemata are acquired in an order of complexity and that solution of problems ordered by complexity can effectively capture differential levels of schema acquisition. To determine if the order of mastery affects the nature of problem solving errors, the following hypothesis, stated in alternative form, is tested:

Hypothesis 3: Errors in problem solving by decision aid users will be dominated by the use of an incorrect, simple template in place of the correct, more complex template.

METHOD

Overview of Experiments

An experimental approach is used to test the hypotheses. Each of the two experiments consisted of three phases: a schema acquisition phase, a distracter phase, and a performance measurement phase. Experimental materials were similar to those used in Rose and Wolfe (2000). In the schema acquisition phase, subjects first completed a set of three practice problems involving the calculation of taxable income and tax liability for single taxpayers with capital gains and social security income. To solve the practice problems, subjects in three treatments used computer-based decision aids, and a fourth treatment group completed them by hand. All subjects then performed a distracter task to clear working memory. In the final phase, the subjects completed a set of test questions consisting of five levels of complexity to measure acquired problem-solving schemata.

The second experiment was identical to the first, with the exception of the final phase. Subjects in the second experiment completed a set of five text-editing questions to measure schema acquisition.[2] Subjects completed the entire experiment during one session in a university computer lab (approximately 90 minutes).

Subjects and Compensation

Subjects consisted of accounting senior volunteers from two large universities. Given that the purpose of the experiment was to measure schema acquisition, student subjects were necessary because participants could not have existing capital gains or social security schemata. Therefore, the subject selection procedure required the exclusion of subjects with existing tax knowledge. In addition, a knowledge pretest was used to remove subjects with tax knowledge. Of the 186 students who volunteered to participate, 10 were excluded from the analysis due to previous tax-related course work, tax experience, or tax knowledge. The 176 subjects analyzed had an average GPA of 2.97, age of 22 years, and 60% were female. All subjects were within one year of graduation and were declared accounting majors. As such, the subjects are reasonable surrogates for entry-level tax professionals that have considerable opportunities to learn through experience.

All volunteers received extra credit in their accounting classes for successful completion of the entire experiment. To motivate students to exert effort during the experiment, they were paid based upon their performance. Subjects received $2.00 for each question answered correctly (up to a maximum of $10). Subjects were aware that performance-based compensation was involved when they started the experiment, and they were told at the start of the experiment that their goals were to arrive at the correct solutions to the practice problems and to learn as much tax material as possible. Subjects were informed that the compensation scheme related to the performance measurement phase of the experiment when they began that phase of the experiment. By having the simultaneous goals of solving the practice problems correctly and learning the tax material, the subjects were exposed to the same conditions that would be expected in practice-based experiential learning.

The Schema Acquisition Phase

The manipulations were performed across four treatments in the schema acquisition phase of the experiment, and subjects were randomly assigned to each of the experimental units. Treatments groups one, two, and three completed three practice problems with decision aids; and each treatment's decision aid was manipulated to induce a different level of cognitive load. Treatment four solved the same three practice problems, but performed all calculations by hand. To begin the schema acquisition phase, subjects in each treatment read a passage from their computer screens which stated that their goals in solving the three practice problems were to learn how to compute tax liabilities and to solve the

problems as accurately as possible. The treatments that used a decision aid could complete the practice problems without attending to the tax calculation instructions, but in order to learn how to calculate tax liabilities, the decision aid users had to attend to the calculation instructions.[3]

Subjects in treatment one received the decision aid that produced the greatest split attention effect. The decision aid for calculating tax liabilities was presented on one screen, and subjects were required to change to another screen to view the tax calculation instructions. This aid was shown by Rose and Wolfe (2000) to induce a heavy cognitive load, which significantly reduced subjects' ability to learn from the aid. Subjects could switch screens at any time and as many times as they wished. Switching screens induces a heavy cognitive load because subjects must take information from physically separate locations and hold it in working memory while integrating the information (Mousavi et al., 1995; Rose & Wolfe, 2000). Subjects in the second treatment used a decision aid with the aid steps and instructions presented on one screen. This aid produced less cognitive load than the aid in treatment one. In the third treatment, subjects received the decision aid with the tax calculation instructions integrated into the steps performed by the aid. This treatment produced the least split attention and, therefore, the lowest cognitive load (Rose & Wolfe, 2000).

Subjects in treatment four solved the practice problems by hand. Consequently, the split attention level in this treatment is not directly comparable to the treatments using a decision aid. The no-aid treatment allows for comparison of learning across users and non-users of decision aids, and the no-aid treatment received the same tax calculation instructions as the aided treatments.

The decision aids were identical with the exception of the placement of tax calculation instructions. Treatment groups were physically separated to prevent subjects from recognizing any differences in treatment. To control for potential media effects, all treatments performed the task on identical computers, including the hand-calculation group. Subjects were required to input their answers for taxable income and tax liability into specified boxes in the decision aid, after which they could receive feedback in the form of the correct answers by clicking on a "CORRECT ANSWER" button. Feedback was necessary to prevent subjects from generating schemata for improper solution strategies. Note that the feedback was not explanatory; explanation was provided in the manipulated computation instructions.

After checking answers, subjects were allowed as much time as they desired to reach the correct solution and study and work the problems further. When subjects were ready to begin a new practice problem, they clicked on the "DONE" button. The computer captured the total time spent on each practice

problem.[4] All subjects were allowed the use of scratch paper and identical calculators throughout this phase.

Performance Measurement

Upon completion of the practice problems, subjects were presented with a screen informing them to turn in their scratch paper and to begin the next phase of the experiment. The first steps in this phase involved the collection of demographic information and a distracter task designed to clear working memory. The purpose of the experiment is to measure schema acquisition, not working memory retention. Therefore, working memory had to be cleared before the performance measurement phase. Subjects were required to perform three subtraction problems in their heads. The demographic questionnaire itself also acted to clear working memory.

The final step in the performance measurement phase of the experiment involved measuring the schemata acquired. Schematic knowledge can be assessed by having subjects: (1) identify information in problems that is necessary and sufficient for solution (Low & Over, 1990); or (2) solve problems analogous to practice problems (Rose & Wolfe, 2000; Low & Over, 1992). This research employed both measures.

For the first measure of schema acquisition, subjects were required to solve five test problems analogous to practice problems. These problems are identical to those used in Rose and Wolfe (2000). One problem was presented for each of the following five taxation topics, which are ordered by level of complexity:

Problem (1) Problem requires knowledge of the concepts of adjusted gross income, deductions, and exemptions.

Problem (2) Problem requires knowledge of the tax liability computation rules for long-term capital gains when taxpayers are in the 15% or 28% tax brackets (long-term capital gains taxed as ordinary income).

Problem (3) Problem requires knowledge of the tax liability computation rules for long-term capital gains when taxpayers are in tax brackets above 28% (long-term capital gains taxed at maximum marginal rate of 28%).

Problem (4) Problem requires knowledge of the tax computation rules for taxable social security benefits when provisional income does not exceed the first base amount.

Problem (5) Problem requires knowledge of the tax computation rules for taxable social security benefits when provisional income exceeds the first base amount.

The five test problems related to the above topics flow from least complex to most complex. Complexity is measured using the number of rules required for problem solution (Sweller, 1988; Low & Over, 1992). The number of rules needed to solve the test problems at each of the five levels of complexity are as follows: level one requires three rules, level two requires five rules, level three requires six rules, level four requires eight rules, and level five requires 11 rules. Following Low and Over (1992), the material inherent in the level one test problem is required knowledge for test problems at all other levels, but the four test problems above level one are independent with respect to required knowledge. To control for potential order effects, test problems were given to subjects in a random order.

Each test problem was scored from 0 to 4 points.[5] The total score on the test problem task can range from 0 to 20 points. Differences in the total score of the five test problems capture differences in the level of schematic knowledge acquired, because the problems are ordered by the level of complexity. Therefore, higher scores correspond to higher-level schematic knowledge.

The second experiment employed a second measure of schema acquisition that required text editing. In the text-editing task, subjects were required to indicate whether a problem provided sufficient, irrelevant, or insufficient information for solution. Actual solution of the problem was not required. In cases where information was insufficient, subjects were required to provide the missing information; and in cases where irrelevant information was present, subjects were required to circle the irrelevant information. Text editing provides an excellent measure of schema acquisition because accurate solution of text editing problems indicates that subjects recognize problems that belong to a specific category and require specific moves for solution. This is precisely the definition of schemata provided by Tarmizi and Sweller (1988). In addition, text editing removes the measurement noise in the analogous test problems created by computation and mathematical errors.

Text editing problems were graded as correct if they were correctly classified as sufficient, irrelevant, or insufficient, and any irrelevant or insufficient information was correctly identified. For the text-editing task, only three levels of complexity could be evaluated, as opposed to the five levels evaluated in the task with the analogous test problems of varying complexity. To design text-editing questions for all five levels of problem complexity was not possible because of the nature of text editing problems. Levels four and five, for example, require the same information units for solution, although different rules are necessary for solution. When problems require different sets of rules, but the same information for solution, text-editing questions will not distinguish between the problems. The levels of complexity examined in the text-editing task were:

Level (1) Problems requiring knowledge of the concepts of adjusted gross income, deductions, and exemptions (same complexity as Problem (1) from the problem solution task).

Level (2) Problems requiring knowledge of the tax liability computation rules for long-term capital gains (includes concepts in Problems (2) and (3) from the problem solution task).

Level (3) Problems requiring knowledge of the tax computation rules for taxable social security benefits (includes concepts in Problems (4) and (5) from the problem solution task).

The text-editing task included one problem at Level 1, two problems at Level 2, and two problems at Level 3. The total score on the text-editing task can range from 0 to 5 points.

RESULTS

Cognitive Load and Schema Acquisition

The descriptive statistics presented in Panel A of Table 1 indicate that the mean scores of the test problems in the performance measurement phase of the experiment increase as cognitive load decreases. Analysis of covariance (ANCOVA) and related mean comparisons were performed and are presented in Panels B and C of Table 1.[6] The ANCOVA model with test problem score as the dependent variable is statistically significant at the 0.001 level. Tests for differences between treatment means were conducted using Student-Newman-Keuls mean separation procedures. The mean separations indicate that subjects using the integrated instructions decision aid scored higher on the test problems than did subjects with instructions on a separate screen, providing support for hypothesis one. The findings indicate that the treatment using the decision aid that produced the lowest cognitive load acquired significantly more complex schemata than the treatment using the decision aid that produced the highest cognitive load. This replicates the results of Rose and Wolfe (2000).

Panel A of Table 2 presents descriptive statistics for text editing scores. Similar to the problem scores in Table 1, the descriptive results indicate that the mean scores on the text editing problems increase as cognitive load decreases. Panel B of Table 2 present the ANOVA results for a model using the text editing problem score as the dependent variable and Panel C presents the related mean comparisons.[7] These results with an alternative measure of schema acquisition again indicate that reducing cognitive load increases schema acquisition. Subjects using the aid with integrated instructions (i.e. low cognitive load) correctly solved

Table 1. Descriptive Statistics and ANCOVA for Test Problems Ordered by Complexity.

Panel A: Descriptive Statistics for Test Problems

Treatments	*N*	Mean[a]	Deviation	Minimum	Maximum
Separate Screen (High Load)	20	6.55	3.46	0	11
Same Screen (Medium Load)	20	7.95	4.35	0	14
Integrated (Low Load)	19	10.05	1.93	8	14
No Aid	22	13.68	3.91	3	19
Total	81	9.65	4.45	0	19

Panel B: ANCOVA for Test Problems Scores

Source	Sum of Square	d.f.	Mean Square	F	Sig.
Treatments	384.318	3	128.106	10.69	0.001
Covariate (Total Time)	62.873	1	62.873	5.247	0.025
Error	910.747	76	11.984		
Total		80			

R square = 0.425.

Panel C: SNK Mean Comparisons for Test Problems Scores

		Differences in Treatment Least Square Means			
Treatment	LS Mean	Separate	Same	Integrated	No Aid
Separate Screen (High Load)	6.99	–	1.22	2.94*	6.16**
Same Screen (Medium Load)	8.21		–	1.72	4.94**
Integrated (Low Load)	9.93			–	3.22**
No Aid	13.15				–

* Significant at $p < 0.05$; ** Significant at $p < 0.01$.

[a] Means represent the total score on the test problem task, where each test problem was scored from 0 to 4 points.

more text editing problems than subjects with the instructions on a separate screen. Unlike the results for the problem solving scores, there is no significant difference between the text editing scores of subjects in the unaided condition and the low cognitive load condition.

Acquisition Order and Complexity of Schematic Knowledge

Hypothesis two predicts that simple problem solving templates, i.e. those based on fewer rules, will be acquired by subjects before more complex templates

Table 2. Descriptive Statistics and ANCOVA for Text Editing Problems.

Panel A: Descriptive Statistics for Text Editing Problems

Treatments	*N*	Mean[a]	Deviation	Minimum	Maximum
Separate Screen (High Load)	33	2.12	1.08	1	4
Integrated (Low Load)	31	2.74	0.97	1	4
No Aid	31	2.84	1.07	1	5
Total	95	2.56	1.04	1	5

Panel B: ANOVA for Text Editing Scores

Source	Sum of Squares	d.f.	Mean Square	F	Sig.
Treatment	9.79	2.00	4.89	4.52	0.01
Error	99.64	92.00	1.08		
Total	95.00				

R square = 0.100.

Panel C: SNK Mean Comparisons for Text Editing Scores

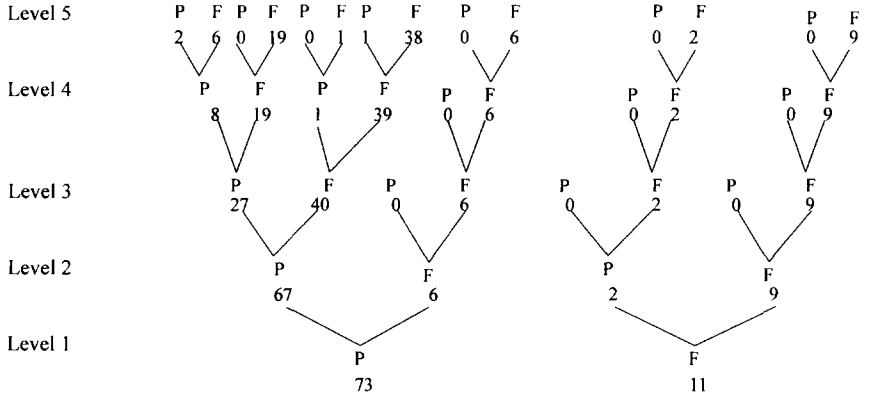
		Differences in Treatment Means		
Treatment	Means	Separate	Integrated	No Aid
Separate Screen (High Load)	2.12	–	0.62*	0.72*
Integrated (Low Load)	2.74	0.10		
No Aid	2.84			

* Significant at $p < 0.05$.

[a] Means represent the total score on the text editing task, where each text editing problem was scored as 0 (incorrect) or 1 (correct).

are acquired. For the first test of this hypothesis, the analogous test problems across five levels of complexity were scored as either correct or incorrect. Figure 1 shows the number of test problems answered correctly and incorrectly at each level of complexity. Most subjects correctly solved the level one test problem, which required comprehension of adjusted gross income, taxable income, and the use of the tax table. The second level problem required knowledge of how to compute tax on capital gains for taxpayers in the 15% and 28% tax brackets. Of the 69 subjects who correctly solved the second level problem, only two of the subjects failed to correctly solve the level one problem.[8] The third level problem required knowledge of the rules for tax calculations when long-term capital gains are not taxed as ordinary income. All subjects who correctly solved the third level problem also correctly solved the problems for levels one and

two. The fourth and fifth level problems required knowledge of the taxation of social security benefits. In the level four problem, provisional income does not exceed a base amount, while in the level five problem the provisional income does exceed a base, and a portion of the social security benefit is taxable. Only one subject who correctly solved the fourth level problem did not solve the level three problem correctly, and one subject who correctly solved the fifth level problem failed on the level four problem.



P = problem passed (answered correctly)
F = problem failed (answered incorrectly)

Level 1) Problems requiring knowledge of the concepts of adjusted gross income, deductions, and exemptions.

Level 2) Problems requiring knowledge of the tax liability computation rules for long-term capital gains when taxpayers are in the 15% or 28% tax brackets.

Level 3) Problems requiring knowledge of the tax liability computation rules for long-term capital gains when taxpayers are in tax brackets above 28%.

Level 4) Problems requiring knowledge of the tax computation rules for taxable social security benefits when provisional income does not exceed the first base amount.

Level 5) Problems requiring knowledge of the tax computation rules for taxable social security benefits when provisional income exceeds the first or second base amount.

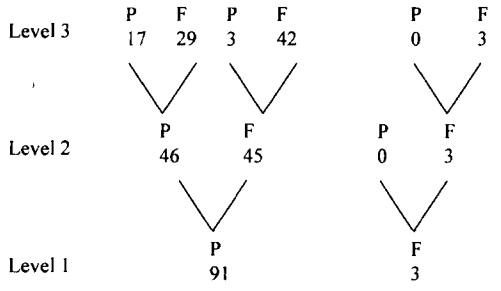
Fig. 1. Tree Diagram of Test Problems Answered Correctly.

Based upon Fig. 1, subjects apparently acquired the tax liability schemata in a linear order of complexity in the experimental task. The tree diagram in Fig. 1 can be analyzed more formally using Guttman scalogram analysis, which furnishes a coefficient of reproducibility that indicates the predictability of the pattern of schema acquisition. According to Guttman analysis, any coefficient greater than 0.90 indicates a very strong pattern, and for the results illustrated in Fig. 1, the coefficient is 0.98 (Maranell, 1974). This finding indicates that the relative order of mastery in this group of tax problems is constant across subjects.[9] In support of the second hypothesis, results indicate that subjects who do not acquire simple problem solving templates also fail to acquire more complex templates. In addition, the evidence for acquisition of schemata in an order of complexity indicates that higher scores on the test problems correspond to the acquisition of more complex schemata.

Acquisition order can also be analyzed using the text editing problems. Figure 2 presents a tree diagram of the accuracy on text editing problems.[10] Again, the tree diagrams indicate that subjects who are unable to solve the less complex text editing questions are also unable to answer the more complex problems. Those subjects who correctly solve the most complex text editing problems also answer the first levels correctly. Guttman scalogram analysis of the tree diagrams yields coefficients of reproducibility of ranging from 0.97 to 0.98. Both measures of schema acquisition strongly support the hypothesis that the tax problem schemata are acquired in a linear order of complexity. The results for text editing also support the effectiveness of using problems ordered by complexity to capture differential levels of schema acquisition. Results are robust across both measures, but problem complexity tasks are easier to implement, can be applied in numerous domains, and allow for additional decomposition of complexity.

To further analyze the efficacy of using analogous problems ordered by complexity to measure schema acquisition, the errors made in solving the test problems were investigated. Hypothesis three predicts that test problem errors will be dominated by the incorrect use of simple templates in place of the correct, complex templates. Problems from levels three, four, and five present the only opportunities to test this hypothesis, because subjects have the ability to mistakenly use inappropriate, simple templates at these levels.[11] Panels A, B, and C of Table 3 provide descriptions of the errors made on problems in levels three, four, and five and their relative frequencies.

The errors of greatest interest involve the application of simple solution templates when more complex templates are appropriate. During solution of level three problems, subjects can incorrectly use level two templates. That is, subjects can treat all long-term capital gains as ordinary income and fail to



P = problem passed (answered correctly)
F = problem failed (answered incorrectly)

Level 1) Problems requiring knowledge of the concepts of adjusted gross income, deductions, and exemptions (same complexity as Problem 1 from the problem solution task).

Level 2) Problems requiring knowledge of the tax liability computation rules for long-term capital gains (includes concepts in Problems 2 and 3 from problem solution task).

Level 3) Problems requiring knowledge of the tax computation rules for taxable social security benefits (includes concepts in Problems 4 and 5 from problem solution task).

Fig. 2. Tree Diagram of Text Editing Problems Answered Correctly.

consider the 28% cap on these gains. Six different types of errors were found on the level three problem. However, of the 65 errors made on this problem, 52 errors resulted from the application of a level two template to the level three problem (Chi Square 188.59, $p < 0.000$). In level four problems, subjects may rely on a level one template and treat all social security benefits as ordinary income. Of the 74 errors made on the level four problem, 35 of these errors involved the application of a level one template to the level four problem. Four other types of errors were made on the level four problem, but again, the use of an inappropriate, simple template was more prevalent than any other form of error (Chi Square 38.973, $p < 0.001$).

Errors of interest on the level five problem are defined either by the improper application of the level five template (taxing an incorrect portion of the social security benefit), by the inappropriate use of the level four template (incorrectly treating the social security benefits as non-taxable), or by the inappropriate use of the level one template (incorrectly treating social security benefits as ordinary income). These errors accounted for 73 of the 81 made on the level five problem: 23 subjects applied the level one template, 18 subjects applied the level four template, and 32 subjects attempted to apply the correct level five template, but did so incorrectly. These results indicate that subjects "fell

Table 3. Analysis of Errors Made on Test Problems.

Panel A: Frequency of Errors Made on the Level Three Test Problem

	Frequency	Percent	Cumulative Percent
Error 1 – Applied level two template	52	80.00%	80.00%
Error 2 – Arithmetic Error	5	7.70%	87.70%
Error 3 – All gains taxed at 98%	1	1.50%	89.20%
Error 4 – LT gains not taxed	2	3.10)%	92.30%
Error 5 – Tax table used incorrectly	2	3.10%	95.40%
Error not identifiable	3	4.60%	100%
Total	65	100%	

Chi-Square = 188.569, $p < 0.000$.

Panel B: Frequency of Errors Made on the Level Four Test Problem

	Frequency	Percent	Cumulative Percent
Error 1 – Applied level one template	35	47.30%	47.30%
Error 2 – Did not calculate provisional income	11	14.86%	62.16%
Error 3 – Provisional income calculated incorrectly	13	17.57%	79.73%
Error 4 – Used provisional income in place of AGI	1	1.35%	81.08%
Error not identifiable	15	20.27%	100%
Total	74	100%	

Chi-Square = 38.973, $p < 0.001$.

Panel C: Frequency of Errors Made on the Level Five Test Problem

	Frequency	Percent	Cumulative Percent
Error 1 – Error within level five template	32	39 50%	39.50%
Error 2 – Applied level four template	18	02.00%	61.70%
Error 3 – Applied level one template	23	28.40%	90.10%
Error not identifiable	8	9.90%	100%
Total	81	100%	

Chi-Square = 14.850, $p < 0.005$.

back" to the simplest template that they had acquired (Chi Square 14.852, $p < 0.005$).

Chi-Square tests that compare expected error frequencies to actual error frequencies for the problems at levels three, four, and five all indicate that the use of incorrect, simple templates occurs significantly more often than any other form of error (see Table 3). These findings support hypothesis three. Subjects

rely on simple problem-solving schemata that have been acquired when faced with problems requiring the application of more complex schemata that have not been acquired.

Finally, a breakdown by treatment of the errors made in the level five-test problem provides additional support for the findings on the effects of cognitive load on schema acquisition in a decision aid learning environment. As noted, the level five problem required comprehension of the taxation of social security benefits when provisional income exceeded a base amount, and all identifiable errors made during the solution of the problem can be traced to the incorrect use of the level five template or the inappropriate application of level one or level four templates.[12] Table 4 displays the frequencies of errors made on the level five test problem, organized by treatment.

Table 4. Analysis of Errors Made on the Level 5 Test Problem.

		Treatment				
		Separate Screen	Same Screen	Integrated	No Aid	Total Errors
Error 1						
	Frequency	10	9	3	1	23
	% of error 1	43.50%	39.10%	13.00%	4.30%	100%
	% within treatment	55.60%	42.90%	18.80%	5.60%	
Error 2						
	Frequency	3	4	5	6	18
	% of error 2	16.70%	22.20%	27.80%	33.30%	100%
	% within treatment	16.70%	19.00%	31.30%	33.30%	
Error 3						
	Frequency	5	8	8	11	32
	% of error 3	15.60%	25.00%	25.00%	34.40%	100%
	% within treatment	27.80%	38.10%	50.00%	61.10%	
Total within treatments:						
Frequency		18	21	16	18	73
%		100%	100%	100%	100%	

Chi-Square = 12.985 ($p < 0.05$).

Error 1 = Subjects used the level one template. Subjects treated all social security income as ordinary.

Error 2 = Subjects used the level four template. Subjects did not tax any of the benefits.

Error 3 = Subjects used the level five template incorrectly. Subjects taxed one-half of the social security benefits when one-half of the excess of provisional income over the base amount should have been taxed.

The first error involves treating the social security benefit as ordinary income, i.e. applying a level one template. This error is the most basic, because it represents falling back to the lowest possible level. Results indicate that the group with the most cognitive load, the separate screen treatment, was the *most* likely to apply a level one template to the level five problem (55.6% of all errors made by subjects in the separate screen treatment). The group with least cognitive load, the integrated instructions screen, was the *least* likely to apply a level one template to the level five problem (18.8% of all errors made by subjects in the integrated instructions treatment). The third error type in problem five is the most advanced, because it represents recognition of the needed level five template, but the inability to correctly apply that template. Here results show that the group with the most cognitive load, the separate screen treatment, was the *least* likely to attempt a level five template on the level five problem (27.8% of all errors made by subjects in the separate screen treatment). Whereas, the group with least cognitive load, the integrated instructions treatment, was the *most* likely to attempt a level five template on the level five problem (50.0% of all errors made by subjects in integrated instructions treatment).

Overall, the level of cognitive load has a statistically significant effect on the distribution of problem five errors (Chi-Square = 12.985, $p < 0.05$). Taken together, the results from the tests of all hypotheses indicate that increasing cognitive load reduces the complexity of schemata acquired, which leads subjects to rely on simpler templates when attempting to solve tax problems requiring more complex templates. Further, error analysis and tests involving an alternative measure of schema acquisition both indicate that test problems ordered by complexity can effectively capture differences in schema acquisition and reveal the level of complexity of schemata that aid users have acquired.

Although no formal hypotheses were developed for the non-aid users, analysis of their learning performance and errors reveals valuable insights into the effects of decision aids on learning relative to traditional learning environments. Tables 1 and 2 indicate that the no-aid group developed more complex schemata than all aid users, although significant differences in text editing performance between the low load aid users and the non-users did not exist. This result is generally consistent with prior research, which has found that decision aid users acquire less declarative and procedural knowledge than non-users (see, e.g. Moffitt, 1994; Pei et al., 1994; Steinbart & Accola, 1994; Odom & Dorr, 1995). The finding that aid users learn less than non-users should not be alarming. If learning maximization was the sole function of decision aids, then perhaps aids should be eliminated from practice. But decision aids offer efficiency, consistency, accuracy, and documentation benefits (Messier, 1995). Given that firms

value these benefits, and that experiential learning is an important component of the accounting profession, low cognitive load aids appear to offer significant benefits relative to high cognitive load aids. Improved decision aid design can maximize the learning that occurs during decision aid use, while preserving the other benefits aids provide. Decision aids are becoming a major component of the development of professional skill, and research into their design is essential.

DISCUSSION

Tax problem schemata are acquired in a linear order from simple to complex when the acquisition stems from computerized decision aid use. This result parallels the findings in educational psychology for non-aided learning environments (Low & Over, 1990, 1992; Sweller, 1994). Decision aid users must be allowed the opportunity to develop simple schemata before more complex schemata will be acquired, even when simple schemata are not requisite knowledge for the acquisition of more complex schemata. Further, decision aid users who acquire only simple schematic knowledge mistakenly use it when complex schemata are needed. These results suggest that individuals trained on a decision aid that produces a high cognitive load would commonly make the mistake of inappropriately applying simple schemata when performing calculations without the benefit of the decision aid.

Use of multiple procedures to measure schema acquisition and the error analyses indicate that problems ordered by complexity are appropriate for capturing differences in schema acquisition. This finding both validates the results of Rose and Wolfe (2000) and presents significant opportunities for further investigation of schema acquisition. Problem complexity tasks can be applied in many domains to measure the effects of system design alternatives and instructional design methods on schema acquisition. This method, as well as text editing tasks, will allow future research to investigate schema acquisition, the foundation of expertise, rather than simply measuring changes in quantities of declarative and procedural knowledge. Problems of ordered complexity and text editing tasks both reveal information about the complexity of knowledge structures, while the simpler test measures used in prior decision aid research captured the quantity of knowledge. Knowledge structure complexity, while not a direct measure of knowledge structure organization, is a much closer approximation of schema acquisition than the knowledge measures employed in the prior literature.

Significant additional opportunities exist to study schema acquisition. The current study examined schema acquisition by novices without existing tax

schemata. Future studies should examine the effect of cognitive load on schema acquisition and refinement by more experienced aid users and in less deterministic task domains. Similarly, the effectiveness of measuring schema acquisition with problems ordered by complexity has not been examined in tasks requiring judgment or with experienced users. Finally, this study finds that tax schemata are acquired in an order of complexity. Other contexts likely exist where learning does not flow strictly from simple to complex schemata. This research does not propose that every learning environment will exhibit the same pattern of schema acquisition.

The error analysis and multiple measures of schema acquisition all confirm the importance of cognitive load to schema acquisition. The cognitive load imposed by a decision aid appears to be a significant factor for system designers to consider. Accounting firms have automated many tasks with decision aids, and these firms expect their accountants to learn from on-the-job experiences, which include the use of decision aids. Cognitive load produced by split attention effects reduces an individual's ability to learn from a decision aid. If novice accountants are able to learn more advanced tasks when using a decision aid, because of decreases in split attention effects produced by the aids, then consideration is warranted for this aspect of decision aid design.

NOTES

1. The order of complexity is based upon Mayer's (1981) classification of algebraic word problems. Complexity is defined by the number and type of propositions related to the problem-solving template needed to solve a particular problem.

2. Subjects who completed the text editing questions also received a short training exercise where they were exposed to the text editing question format. The training exercise involved the solution of geometry problems, which is unrelated to the experimental task. Pilot testing indicated that subjects were not familiar with text editing, and training eliminated confusion regarding the novel question format. Subjects who completed the text editing problems also solved the same five problems as the remaining subjects. The results for these problems are not used in the analyses because they could be influenced by the text editing task. Subjects were required to solve the problems such that the pay scheme would be identical for both treatment groups.

3. The computer program used to administer the experiment captured attention to instructions. Analysis of the program indicated that subjects did attend to the instruction screen.

4. This process of learning while problem solving defines experiential learning, and the measure of time spent on practice problems represents the effort duration related to experiential learning. Subjects were allowed unlimited time for solving the practice problems to prevent ceiling effects resulting from of individual differences in processing speeds.

5. There was no need to compute the interrater reliability as the grading scale was completely objective. The problem grader was unaware of the treatments associated with

each problem set. Additionally, analyses for all treatment differences were repeated using scores of 0 (incorrect) and 1 (correct) for the test problems. No qualitative differences in results were found.

6. Time spent on the first phase of the experiment (the schema acquisition phase) is included in the model to control for variations in effort duration. Some subjects input little effort into the schema acquisition phase and, as a result, acquired less complex schemata. Removal of the effort duration covariate does not cause any substantive changes in results.

7. Effort duration was not statistically significant in the text editing model and was not included in the final analyses.

8. Note that it is not possible to answer the level two problem correctly without the knowledge required to solve the level one problem. It is likely that these two students made careless errors.

9. This finding also validates the measure of complexity, the number of rules needed to solve a problem. As shown in Fig. 1, as problems increased in complexity they were less likely to be answered correctly.

10. There were five text editing problems. One problem was included for the first level of difficulty, two problems for the second level of difficulty, and two problems for the third level of difficulty. The combinations of the five text editing problems result in four possible tree diagrams (i.e. one diagram for each combination of a level 1, level 2 and level 3 problem. While only one tree diagram is presented in Fig. 2, all four combinations produce highly similar results with similar Guttman scores.

11. Level two problem errors cannot be distinguished from level one problem errors, because taxpayers in the 15% and 28% brackets have long term capital gains taxed as ordinary income.

12. The level five problem is the only one which offers such a rich array of relevant errors. The level four problem dealt with non-taxable social security benefits. Therefore, subjects did not incorrectly use capital gains templates, i.e. level two or three templates. The level three problem had no errors involving the incorrect use of the correct template.

REFERENCES

Anderson, J. R. (1985). Role of reader's schema in comprehension, learning, and memory. In: H. Singer (Ed.), *Theoretical Models and Processes of Reading*. Newark, DE: International Reading Association.

Baddeley, A. D. (1992). Working memory: The interface between memory and cognition. *Journal of Cognitive Neuroscience, 4*, 281–288.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*, 293–332.

Chandler, P., & Sweller, J. (1992). The split attention effect as a factor in the design of instruction. *British Journal of Educational Psychology, 62*, 233–246.

Chase, W. G., & Simon, H. (1973). Perception in chess. *Cognitive Psychology, 1*, 55–81.

Chi, M., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In: R. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence*. Hillsdale, NJ: Erlbaum.

Gick, M. (1986). Problem-solving strategies. *Educational Psychologist, 21*(1–2), 99–120.

Glover, S., Prawitt, D., & Spilker, B. (1997). The influence of decision aids on user behavior: Implications for knowledge acquisition and inappropriate reliance. *Organizational Behavior and Human Decision Processes, 72*, 232–255.

Libby, R. (1985). Availability and the generation of hypotheses in analytic review. *Journal of Accounting Research, 23*, 648–667.

Low, R., & Over, R. (1990). Text editing of algebraic word problems. *Australian Journal of Psychology, 43*, 63–70.

Low, R., & Over, R. (1992). Hierarchical ordering of schematic knowledge related to the area-of-rectangle problem. *Journal of Educational Psychology, 84*, 62–69.

Maranell, G. (1974). *Scaling: A Sourcebook for Behavioral Scientists*. Chicago, IL: Aldine Publishing.

Mayer, R. E. (1979). Qualitatively different encoding strategies for linear reasoning premises: Evidence for single association and distance theories. *Journal of Experimental Psychology: Human Learning and Memory*, (Jan.), 1–10.

Mayer, R. E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. *Instructional Science, 74*, 199–216.

Messier, W. (1995). Research in and development of audit decision aids. In: R. H. Ashton & A. H. Ashton (Eds), *Judgment and Decision Making in Accounting and Auditing*. Cambridge University Press.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Moffitt, K. (1994). An Analysis of the Pedagogical Effects of Expert System Use in the Classroom. *Decision Sciences, 25*, 445–457.

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing Cognitive Load by Mixing Auditory and Visual Presentation Modes. *Journal of Educational Psychology, 87*, 319–334.

Odom, M. D., & Dorr, P. (1995). The impact of elaboration-based expert system interfaces on de-skilling: An epistemological issue. *Journal of Information Systems, 9*, 1–18.

Owen, E., & Sweller, J. (1985). What do students learn while solving mathematics problems? *Journal of Educational Psychology, 77*, 272–284.

Pei, B., Steinbart, P. J., & Reneau, J. H. (1994). The effects of judgment strategy and prompting on using rule-based expert systems for knowledge transfer. *Journal of Information Systems, 8*, 21–42.

Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory and Cognition, 17*(4), 398–422.

Rose, J., & Wolfe, C. (2000). The effects of system design alternatives on the acquisition of tax knowledge from a computerized tax decision aid. *Accounting, Organizations, and Society, 25*, 285–306.

Simon, H. (1974). How big is a chunk? *Science, 183*, 482–488.

Simon, H. (1976). Discussion: Cognition and social behavior. In: J. S. Carroll & J. W. Payne (Eds), *Cognition and Social Behaviour* (pp. 258–267). Hillsdale: Earlbaum.

Simon, H., & Gillmartin, K. (1973). A simulation of memory for chess positions. *Cognitive Psychology, 5*, 29–46.

Steinbart, P. J., & Accola, W. L. (1994). The relative effectiveness of alternative explanation formats and user involvement on knowledge transfer from expert systems. *Journal of Information Systems, 8*(Spring), 1–17.

Sweller, J. (1988). Cognitive load during problem solving. *Cognitive Science* 12: 257–285.

Sweller, J. (1993). Some cognitive processes and their consequences for the organisation and presentation of information. *Australian Journal of Psychology, 45*, 1–8.

Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction, 4*, 295–312.

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load and selective attention as factors in the structuring of technical material. *Journal of Experimental Psychology, 119*, 176–192.

Sweller, J., & Levine, M. (1982). Effects of goal specificity on means-end analysis and learning. *Journal of Experimental Psychology; Learning, Memory, and Cognition, 8*, 463–474.

Sweller, J., Mawer, R., & Ward, M. (1983). Development of expertise in mathematical problems solving. *Journal of Experimental Psychology, 112*, 639–661.

Tarmizi, R., & Sweller, J. (1988). Guidance during mathematical problem solving. *Journal of Educational Psychology, 80*, 424–436.

Weber, R. (1980). Some characteristics of the free recall of computer controls by EDP auditors. *Journal of Accounting Research, 18*, 214–241.

MORALITY VS. IDEOLOGY: IMPLICATIONS FOR ACCOUNTING ETHICS RESEARCH

Dann G. Fisher and John T. Sweeney

ABSTRACT

The explosion of accounting ethics research in the past decade has been fuelled by the availability of the Defining Issues Test (DIT: Rest, 1979, 1993). Despite existing criticisms regarding the DIT (Emler et al., 1983) and research indicating that the measure is a biased reflection of accountants' moral reasoning (Sweeney & Fisher, 1998, 1999), the validity of the instrument is rarely questioned. This paper discusses the evolution of the political attitude-moral reasoning debate. A between-subjects experiment provides further support for the existence of political bias, implying that ethics researchers may have confounded political ideology with moral reasoning when interpreting results.

INTRODUCTION

Until the past decade empirical ethics research was a relative footnote in the annals of academic accounting literature. Limited most by the lack of a reliable and valid instrument for measuring individual moral judgment or reasoning ability, this problem was apparently resolved with the arrival of the Defining Issues Test (DIT: Rest, 1979, 1993). Grounded in Kohlberg's (1969) cognitive

Advances in Accounting Behavioral Research, Volume 5, pages 141–160.

ISBN: 0-7623-0953-9

theory of moral development, the DIT acquired credibility from its use in hundreds of studies in psychology and the social sciences (Rest, 1986; Rest et al., 1999a). Inspired by the availability of an objective, pencil-and-paper instrument requiring no special training, behavioral accounting researchers seized the opportunity and DIT-based ethics research occupied a significant portion of journal space by the early 1990s (for a review, see Louwers et al., 1997; Ponemon & Gabhart, 1994).

Yet despite the relatively modest relationships found between the moral reasoning capabilities of accountants, as measured by the *P* score of the DIT, and professional judgment and behavior (Ponemon & Gabhart, 1994), accounting researchers have rarely questioned the validity of the DIT. A long running debate in the psychology literature, however, suggests that the DIT is biased in favor of persons who prefer political liberalism to political conservatism (Emler et al., 1983; Markoulis, 1989; Sparks & Durkin, 1987). Emler et al. (1983) went so far as to assert that the DIT *P* score is effectively a measure of political attitude. Given the predominance of conservatives in the accounting profession (Sweeney, 1995; Eynon et al., 1997; Thorne et al., 2000), the debate infers that the moral reasoning abilities of accountants are understated when measured by the DIT and that accounting ethics researchers may have misinterpreted the results of their studies.[1]

Rest (1986) and his colleagues (Barnett et al., 1995) were critical of the research methodology of Emler et al. (1983), claiming that the findings were an artefact caused by asking subjects to respond to the DIT from an assigned political perspective. Rest (1986) contended that political ideology explained only a small amount of the variance in DIT *P* scores, and dismissed the higher scores attained by persons who labeled themselves as liberal as a developmental difference. Recently, Rest et al. (1999a) admitted that the relationship between political ideology and DIT *P* scores is stronger than initially believed, and conceded that political ideology explained as much as 40% of the variance in *P* scores. Although he acknowledged the strong correlation with political ideology, Rest (Rest et al., 1999a; Thoma et al., 1999) maintained that the *P* score provides significant information beyond that which can be obtained from knowing a person's political ideology. Rest et al. (1999a), however, do not address whether or not the significant overlap between political ideology and DIT *P* scores results in an unduly noisy measure of the moral reasoning construct incapable of being unambiguously interpreted.

Accounting ethics researchers have begun to examine this issue, concluding that the DIT produces a biased measure of an accountant's moral reasoning ability (Fisher & Sweeney, 1998; Sweeney & Fisher, 1998, 1999). Their results indicate that: (1) higher-order DIT response items are associated with political

liberalism; (2) DIT scores can be significantly increased for conservative, moderate and even liberal test-takers by responding from the perspective of an 'extreme liberal'; (3) DIT *P* scores, produced under the standard test instructions, overstate the developmental maturity of liberal accountants and understate the development of conservative and moderate accountants; and (4) much of the bias in DIT scores can be removed by appropriately modifying the test instructions. Based on these results, they concluded that failure to control for subjects' political ideology may confound the relationship between DIT *P* scores and professional judgment, behavior and attitudes (Sweeney & Fisher, 1998, 1999).

The findings of the earlier studies examining the relationship between DIT scores and political ideology (Sweeney & Fisher, 1998, 1999) are limited by the use of a within-subject design. Because they had already completed the DIT using the standard instructions, the subjects in these studies might have been influenced by the new instructions to believe that they should answer differently, thereby artificially changing their scores. The purpose of this paper is to address the methodological limitation of earlier studies by using a between-subjects design and to discuss the implications of failing to control for political ideology in accounting DIT ethics research.

The paper proceeds in the following manner. The evolution of the debate regarding the influence of political ideology on DIT *P* scores is first examined. In addition, empirical evidence indicating that this influence may exist is reviewed. Next, the results of a between-subjects experimental study are presented. These results provide further evidence that political ideology is systematically influencing DIT *P* scores and, taken in combination with those of Sweeney and Fisher (1998, 1999), suggest that relationships between DIT *P* scores and other variables previously attributed to differences in subjects' moral reasoning abilities may, in reality, have been largely the result of variance in political ideologies. Finally, a sample of accounting ethics studies that employed the DIT as a measure of moral reasoning ability but failed to control for political ideology are re-interpreted in light of these findings. Alternative explanations for the relationships found between DIT *P* scores and the variables of interest are provided for these studies. The final section consists of our discussion and conclusions.

EVOLUTION OF THE POLITICAL BIAS DEBATE

Kohlberg's (1969, 1981) moral development theory maintains that people use three general hierarchical approaches to resolving moral dilemmas, which he classifies into levels: pre-conventional, conventional, and post-conventional or

principled. Each level is further divided into two stages, with the second stage representing a more advanced perspective of the general strategy. The six stages constitute a developmental sequence of moral problem solving strategies, with the simpler stages preceding increasingly more complex stages.

The theory of moral development is grounded in traditional liberal philosophical thought (Kohlberg, 1969, 1981), leading Kohlberg to identify post-conventional reasoning with political liberalism (Kohlberg 1981, 1983).[2] Empirical research (Emler & Stace, 1999; Fishkin et al., 1973; Fontana & Noel, 1973; Haan et al., 1968; Nassi et al., 1983; Rest, 1976) supported this position while linking conservative positions with conventional reasoning. Liberal enlightenment was argued to provide greater sophistication in the general principles for organizing society, allowing the person to transition from the "law and order" constraints of conventional reasoning (Kohlberg, 1981).

The Defining Issues Test

The most commonly used measure of moral reasoning ability is the Defining Issues Test, developed by James Rest (1979, 1986, 1993), a student of Kohlberg. Designed to be substantially consistent with Kohlberg's (1969) justice-based notion of morality, the DIT requires a subject to select prototypic responses that are aligned with his or her point of view on specific hypothetical dilemmas. The long form of the DIT consists of six dilemmas, while the short form utilizes three dilemmas. For each dilemma, twelve issue statements are provided. These statements are designed to represent the manner in which crucial issues might be conceived from the perspective of different stages of reasoning. Using a five-point scale, the test-taker rates the importance of each of the 12 issue statements in determining what ought to be done in that situation. After rating the 12 issue statements for each dilemma, the subject is asked to rank the four issue statements considered to be the most important overall.

Using the rankings, scores may be obtained for each stage of reasoning (Stages 2 through 6). The reliability of the individual stage scores, however, has not been as high as that for an index of principled reasoning – the *P* score (Rest, 1979, 1993). The *P* score, reported as a percentage of a subject's rankings for Stages 5 and 6, can range from 0 to 95. Rest (1993) reports an average *P* score of 40 for the general population of American adults and an average *P* score of 45 for college graduates. The *P* score for the short version correlates highly (approximately 0.90) with the *P* score for the long version (Rest, 1993).

In addition to the Stage items, the DIT contains "*A*" items designed to represent an anti-establishment orientation. The *A* items condemn tradition and established social order for its arbitrariness, inconsistency, and exploitation

without offering a positive alternative (Rest, 1979). The *A* score stems from a study by Kohlberg and Kramer (1969) that found that college students appeared to regress to Stage 2 reasoning during the transition between conventional and principled moral judgment. As a result, this stage was given a label of $4^1/_2$.

Criterion for a Valid Measure of Moral Reasoning Ability

The *P* score of the DIT has been subjected to numerous studies examining its reliability and validity (Rest, 1979, 1993). Our specific concern is with the construct validity of the *P* score (Cook & Campbell, 1979). Based upon the results of this study and those of previous research (Fisher & Sweeney, 1998; Sweeney & Fisher, 1998, 1999), the *P* score appears to confound moral reasoning with political ideology. The following criteria are critical to the validity of a moral reasoning measure.

The Ceiling Effect

A basic tenet of moral development theory is that person's progress through the stages in a sequential manner (Kohlberg, 1969; Rest, 1979). A person can comprehend the logic of stages he or she previously passed through, but is cognitively incapable of understanding moral reasoning more advanced than his or her developmental level. Therefore, a person can intentionally lower his or her score on a measure of moral reasoning by identifying lower order responses, but should not be able to identify responses representing thinking more advanced than his or her own cognitive capacity. In effect, a person's developmental level establishes a ceiling on his or her understanding of higher order moral reasoning.

Preference vs. Recognition

Rest (1994, p. 17) claims, "higher stages of thinking are preferred until the stage in turn becomes replaced by a newly comprehended stage." As a result, the DIT is reduced to a simple recognition task (Rest et al., 1997): subjects prefer higher order moral reasoning to lower order reasoning and therefore choose the highest order response items recognized (recognition = preference). If the response items of the DIT have an underlying political content, however, then this bias can influence a test-taker's ranking of those items. Consequently, the items an individual prefers to rank as important in resolving a dilemma may not be the same as the items he or she recognizes as representing more advanced thinking. As a result, the *P* score would over- or understate a person's true capacity for moral reasoning.

How might political ideology influence DIT *P* scores? Rest (1994, p. 17) claims that DIT *P* scores produced under standard test conditions represent a

person's highest conceptions of justice and fairness. An unbiased measure elicits a person's highest capacity for moral reasoning because he or she prefers to rank as important those items recognized as representing higher order reasoning consistent with his or her developmental level. But if a politically conservative person comprehends the cognitive complexity of principled DIT responses and chooses to avoid ranking those responses as important because he or she associates this viewpoint with liberalism, then the *P* score would not be measuring this person's most advanced moral thinking. Rest et al. (1999a, p. 142) allude to this possibility:

> Presumably if a subject does not get the point of a DIT item (or does recognize the point but rejects it), the subject passes over the item and goes on to select another item as important.

Similarly, a politically liberal test-taker may overstate his or her DIT *P* score by ranking higher-order response items as important because of their association with liberal ideology, without comprehending the underlying moral content. A valid measure, however, produces an unbiased metric representing the test-taker's true capacity for advanced moral thinking. Such a measure should be free of extraneous influences, such as political associations, that may cause a person to consciously or unconsciously reject more advanced responses although he or she understands the underlying moral reasoning. Conversely, a valid measure of moral judgment will be free of biases that may induce a person to embrace a response reflecting moral thinking more advanced than his or her own cognitive capacity.

The Political Debate

For many years, Rest (1979, 1986, 1993) maintained that political ideology explained only a small amount of the variance in DIT *P* scores, dismissing the higher scores attained by persons who label themselves 'liberal' as a developmental difference. According to Rest (1993, p. 28), correlations between the DIT *P* score and political ideology were either non-significant or low. Critics of moral development theory and the use of the DIT as its primary measurement instrument contended that the DIT was nothing more than a measure of political preference (Emler et al., 1983; Markoulis, 1989).

In order to test the political bias hypothesis, Emler et al. (1983) designed an experiment where subjects first completed the DIT using the standard instructions and then responded from an assigned political perspective, either right-wing, moderate, or left-wing radical. Their results indicated that persons who self-described themselves as right-wingers or moderates generally attained

relatively low *P* scores when completing the DIT in accordance with the standard instructions. However, these individuals were able to significantly increase their scores when asked to respond to the DIT from the perspective of a left-wing radical. This result is inconsistent with a major premise of moral development theory: persons should not be able to fake upward on the DIT because they lack the capacity to comprehend higher order moral arguments (Rest, 1979, 1986).

Rest (1986) claimed that the most damaging result in Emler et al. (1983) to the validity of the DIT, the increase in the mean *P* score, was an artefact of the "radical left-wing" experimental manipulation. By instructing subjects to respond from a radical perspective, Emler et al encouraged them to seek angry sounding items rather than to resolve a moral problem from a politically liberal perspective. Rest maintained that this procedure resulted in subjects rejecting the Stage 4 issue statements that they originally endorsed under the standard DIT instructions and to instead select the anti-establishment (*A*) items. After exhausting the A statements and eliminating Stage 2 and 3 statements as inadequate, Rest asserted that the subjects ". . . had no place left to go but to *P* items" (Rest, 1986, p. 154).

Barnett et al. (1995) provide support for Rest's assertion. In replicating the Emler et al. (1983) study, Barnett et al added additional *A* items to the DIT so that an equal probability of selecting either *A* or *P* items was created. The results indicated that, when responding to the DIT from the left-wing radical perspective, subjects strongly increased their *A* scores and, as a result, decreased their *P* scores below the levels observed using the standard instructions. The results of Barnett et al. (1995) provide an alternative explanation for the findings of Emler et al. (1983) on methodological grounds, but the role of political ideology in affecting DIT *P* scores remained unanswered.

Sweeney and Fisher (1998, 1999) attempted to correct the methodological shortcomings of Emler et al. (1983) by conducting two separate experiments. In the first experiment, subjects (accounting students) first completed the DIT under the standard test instructions. After a two-week period, participants were randomly assigned to respond to the DIT from either an "extremely conservative" or an "extremely liberal" perspective (replacing the "right-wing conservative" and "left-wing radical" terms of Emler et al., 1983). Consistent with Emler et al. (1983), the results of the first experiment indicated that subjects responding to the DIT from an extremely liberal perspective significantly increased their mean *P* scores beyond the levels attained under the standard test conditions.[3] These results confirmed that DIT items were politically loaded but did not establish that test-takers were influenced by this bias in their ranking of the items.

Rest et al. (1999a) are critical of studies such as Sweeney and Fisher (1998, 1999), arguing that manipulation of test instructions should not be used to examine the discriminant validity of the DIT because the instrument and the instructions are "an integrated whole" (Rest et al., 1999a, p. 63). As such, "manipulation of test instructions does not illuminate how people naturally or usually perform a task" (Rest et al., 1997, p. 20). This suggests that the standard DIT instructions prepare subjects to make moral judgments in their usual or normal way – a manner that is purported to elicit their best notions of justice and fairness (Rest, 1979, 1986, 1993). Rest and his colleagues, however, provide no evidence that the standard instructions do in fact elicit the normal or usual manner of reasoning. If the DIT is designed to elicit schema from long-term memory to help sort out what is currently being deliberated, then understanding whether or not the instructions create a setting free of the influence of potential systematic biases, such as political ideology, is critical.

In the second experiment, Sweeney and Fisher (1998, 1999) tested the proposition that the wording of the standard DIT instructions may be a cause of the political bias. The standard instructions direct test-takers to provide their "opinions about social problems". This verbiage may lead subjects to pursue DIT statements consistent with the political image they wish to represent, rather than to seek items they perceive as representing the highest conceptions of justice or fairness. In this experiment, a second sample of accounting students first completed the DIT using the standard instructions. Two weeks later, Sweeney and Fisher again administered the DIT, but instead of using the standard test instructions, the subjects were asked to "score as high as possible" by "identify(ing) the statements designed to represent the highest levels of moral judgment." There should be no difference in *P* scores between the first administration and the experimental condition if the DIT, taken under the standard test instructions, was eliciting a subject's highest capacity for moral reasoning (preference = recognition). Partitioning subjects by political ideology revealed that the mean DIT *P* scores for conservatives and moderates were significantly higher in the experimental condition, where they attempted to score as high as possible, than in the standard test condition. Sweeney and Fisher concluded that the political bias of the DIT resulted in conservative and moderate subjects rejecting items they recognized as higher-order thinking. Liberal subjects actually decreased their *P* scores in the experimental condition. Sweeney and Fisher speculated that under the standard DIT test conditions, liberal test-takers might choose items aligned with their political preferences without recognizing the underlying advanced moral reasoning.[4]

HYPOTHESES

The results of Sweeney and Fisher (1998, 1999) support the argument that the political bias of the DIT influences test-takers ranking of influential items. A potential weakness of their study, however, is the use of a within-subject design. An argument can be made that, given new instructions, subjects simply sought out issue statements different from those they selected in completing the DIT under standard test conditions, thereby artificially changing their scores (Rest, 1986). The objective of this experiment is to methodologically extend the results of Sweeney and Fisher (1998, 1999) by using a between-subjects design to examine whether DIT test instructions are eliciting a test-taker's highest conceptions of justice and fairness.

Rest et al. (1997, p. 20) claim that the standard DIT instructions prepare subjects to make moral judgments in their usual or normal way – a manner which is purported to elicit their best notions of justice and fairness (preference = recognition) (Rest, 1979, 1986, 1993). If the standard instructions, however, prompt test-takers to be influenced by the political content of DIT response items, then a person's recognition of higher-order issue statements would not be equivalent to his or her preference for ranking these items. Politically conservative and moderate test takers may comprehend the cognitive complexity of post-conventional DIT issue statements but avoid ranking these items as important because they associate these statements with liberalism. Consequently, political conservatives and moderates taking the DIT under the standard test instructions should attain lower *P* scores than conservatives and moderates directed to rank issue statements they recognize as representing higher order moral reasoning.

H1: Under the standard instruction set, DIT *P* scores of politically conservative or moderate persons will understate their capacity for advanced moral reasoning.

Conversely, if liberals completing the DIT under the standard test conditions are attracted to issue statements because they reflect a liberal ideology without fully understanding their moral content or if they avoid issue statements that they associate with conservatism, the resulting *P* score will overstate their cognitive capacity for advanced reasoning. As a result, the DIT *P* scores attained by political liberals under the standard instructions should be higher than those of liberals directed to rank issue statements they recognize as representing advanced moral reasoning.

H2: Under the standard instruction set, DIT *P* scores of politically liberal persons will overstate their capacity for advanced moral reasoning.

METHODS

Procedure

A between-subjects design was utilized in order to prevent participants from anchoring on responses recalled from an earlier DIT administration, a potential criticism of the within-subjects methodology of Sweeney and Fisher (1998, 1999). The six-dilemma DIT was administered once to all participants. Subjects were randomly assigned to either the control group or the experimental group. Subjects in the control group were required to complete the DIT under the standard test instructions. In the experimental group, subjects received modified instructions that were consistent with the experimental instructions employed by Sweeney and Fisher (1998, 1999), informing them that:

> The Defining Issues Test is a standardized measure of moral judgment. We are interested in whether you can identify the statements designed to represent the highest levels of moral judgment.

Subjects

A total of 234 junior-level undergraduate accounting students from two midwestern universities voluntarily participated in the study and were randomly assigned to either the control or the experimental condition. The standard instructions for the DIT were provided to the 113 subjects assigned to the control group. The responses of two subjects were eliminated for violating the *M* score manipulation [*M* items are "nonsense items used to check the validity of individuals' responses" (Rest, 1986, p. 42)], while the responses of five subjects were eliminated for failing the consistency check, leaving 106 usable responses (43% male, 57% female). The modified instructions for the DIT were provided to the 121 subjects in the experimental group. The response of one subject was eliminated for violating the *M* score manipulation, while the responses of five subjects were eliminated for failing the consistency check, leaving 115 usable responses (46% male, 54% female).

RESULTS

Subjects were classified into one of three groups – liberal, moderate or conservative – based upon their self-defined political ideology. Consistent with prior research (Emler et al., 1983, Barnett et al., 1995, Sweeney & Fisher, 1999), a seven-point likert scale, with "extremely liberal" and "extremely conservative" as endpoints, was utilized to measure political identity. Subjects responding

from one to three were classified as liberal. Subjects responding in the middle of the scale (four) were classified as moderate. Subjects responding from five to seven were classified as conservative. Reliance on self-defined political orientation is consistent with previous research (Emler et al., 1983, Barnett et al., 1995, Sweeney & Fisher, 1998, 1999) and seems most appropriate since it is one's personal definition that is most likely to influence his or her self-presentation strategy when completing the DIT. For the control group, the described procedure resulted in 29 subjects (27%) being classified as liberal, 33 subjects (31%) being classified as moderate, and 44 subjects (42%) being classified as conservative. For the experimental group, the described procedure resulted in 25 subjects (22%) being classified as liberal, 27 subjects (23%) being classified as moderate, and 63 subjects (55%) being classified as liberal.

The mean *P* score (36.24) of the control group using the standard DIT instructions is consistent with the mean *P* scores of accounting undergraduate students in earlier studies (Ponemon and Gabhart 1994; Sweeney and Fisher 1998, 1999). Moreover, the mean *P* score (38.42) for the group using the modified instructions did not differ significantly ($p = 0.154$) from the mean *P* score (36.24) of the control group using the standard instructions. In isolation, this result could lead one to agree with Rest's (1979, 1986) assertion that the DIT, taken under standard test conditions, yields results that reasonably approximate a person's highest level of moral reasoning. As shown in Table 1, however, a much different result is revealed once political ideology is considered.

Analysis of variance tests generally support H1. For conservatives, the mean DIT *P* score was significantly higher under the modified instructions, directing them to choose items they recognize as reflecting higher order moral thinking, than under the standard instructions. This result is consistent with Sweeney and Fisher (1998; 1999) and suggests that the DIT systematically understates the moral reasoning abilities of political conservatives. Although the mean DIT *P* score of political moderates was higher under the modified instructions than under the standard instructions, the difference was not significant.

As hypothesized in H2, the mean DIT *P* score for liberals was significantly lower under the modified instructions than under the standard test instructions. Liberal subjects attempting to score as high as possible by identifying response statements they recognize as representing advanced moral judgment actually scored lower than liberal subjects responding under the standard instructions. This result is consistent with that of Sweeney and Fisher (1998, 1999), who also found that political liberals attained lower DIT *P* scores under the modified instructions than under the standard instructions. The results are consistent with the contention that the DIT, taken under the standard test instructions, may systematically overstate the moral reasoning capabilities of political liberals.

Table 1. Comparison of DIT *P* Scores – Standard Test Instructions vs. Modified Test Instructions.

LIBERALS	Standard Instructions $n = 29$	Modified Instructions $n = 25$
Mean	43.56	38.93*
Median	43.33	36.67
Std. Dev.	9.88	12.82
MODERATES	Standard Instructions $n = 33$	Modified Instructions $n = 27$
Mean	33.23	37.59
Median	35.00	36.67
Std. Dev.	11.51	12.08
CONSERVATIVES	Standard Instructions $n = 44$	Modified Instructions $n = 63$
Mean	33.67	38.57**
Median	33.33	38.33
Std. Dev.	11.16	10.66

In an analysis of variance, the score using standard instructions differs from the score using the new instructions: * $p < 0.10$; ** $p < 0.05$.

Liberals may choose to rank some post-conventional items because they associate the underlying argument with political liberalism, and not because they understand the underlying moral content.

The results of this experiment suggest that the political bias in the DIT may be driven by the test instructions. When the standard instructions were used, the mean *P* score for liberals is significantly higher ($p < 0.001$) than the mean *P* scores for moderates and conservatives. When the modified instructions were used, the mean *P* scores did not differ by political ideology ($p = 0.920$). Consistent with Sweeney and Fisher (1999), this result suggests that the DIT instructions may be causing subjects to pursue DIT statements consistent with their preferred political ideology, preventing the instrument from presenting a true measure of the person's moral competence.

RE-INTERPRETATION OF ACCOUNTING ETHICS STUDIES

The results of this experiment and those of prior studies (Emler et al., 1983; Sweeney & Fisher, 1998, 1999) suggest that researchers may be unable to

determine how much of the association between the DIT *P* score and the variables of interest is the result of moral developmental differences and how much is attributable to rankings reflecting political ideology. This section demonstrates the potential for misinterpretation by re-examining the results of three selection/socialization studies – Ponemon (1992), Jeffrey and Weatherholt (1996), and Kite et al. (1996) – and providing alternative explanations for the relationships found between DIT *P* scores and the variables examined.

In his widely cited paper, Ponemon (1992) examined the influence of firm socialization on the observed levels of moral reasoning among CPAs. Cross-sectional data indicated that the *P* scores of managers and partners were significantly lower and more homogenous than the *P* scores observed for staff through supervisor levels. This finding was corroborated by longitudinal data revealing that seniors who were promoted to manager had lower and more homogenous DIT *P* scores than those of seniors who chose to leave the firm. Ponemon (1992, p. 243) concluded that

> employees who get a promotion are likely to be perceived by management as having personal characteristics commensurate with the culture and philosophy of the organization.

He further supported his conclusion with survey data showing that managers favored promoting seniors with DIT *P* scores similar to their own, a result that led Ponemon to claim that the critical selection/socialization criterion is a shared set of ethical values.

Although shared ethical values may be an important component of selection/socialization processes, differences in political ideology provide an equally compelling interpretation. Sweeney (1995) found managers and partners in CPA firms to be predominately conservative (85%). If management prefers to promote persons with characteristics similar to their own, then they would likely choose to promote seniors who share their conservative political ideology. As the DIT systematically understates the moral reasoning abilities of conservatives under the standard test instructions, the low DIT *P* scores observed for managers and partners may be explained by a management tendency to promote subordinates with a shared set of conservative political values. Because Ponemon did not control for political ideology, the impact of ethical values and political ideology on promotion decisions in public accounting firms cannot be separated.

Jeffrey and Weatherholt (1996) posited a relationship between an accountant's DIT *P* scores and his or her level of professional commitment, which they defined as the internalization of professional standards. They asserted that high professional commitment would be associated with conventional moral reasoning (Stages 3 and 4), because conventional reasoning is characterized by rule conformance and maintenance of the social order (Rest, 1979). Jeffrey and

Weatherholt found an inverse relationship between DIT *P* scores and level of professional commitment – the mean DIT *P* score for accountants with high professional commitment was significantly lower than the mean DIT *P* score for accountants with low professional commitment. They concluded that the selection/socialization process being observed in practice was the result of accountants with high DIT *P* scores choosing to leave the profession because they are less committed to its rules and values.

The results of Jeffrery and Weatherholt did not address whether the association between DIT *P* scores and professional commitment is the result of differences in moral development or political ideology. One might argue that because the profession is predominately conservative (Sweeney, 1995), conservative accountants would be most likely to embrace the values of the profession. Liberal accountants, on the other hand, might be unwilling to accept the conservative culture of the firm, resulting in alienation toward the profession. In other words, the significant inverse relationship between DIT *P* scores and level of professional commitment observed by Jeffrey and Weatherholt may actually reflect a stronger sense of professional commitment by conservative accountants than by liberal accountants.

This proposition was tested in Sweeney et al. (2001), who controlled for the influence of political ideology on DIT *P* scores in their socialization model of auditors' professional commitment. Their results indicated a strong and significant path between political ideology and professional commitment. After controlling for the influence of political ideology, the path between DIT *P* scores and professional commitment was insignificant. Political ideology, not differences in moral reasoning ability, was an important socialization factor predicting professional commitment in public accounting, calling into question the findings of Jeffrey and Weatherholt (1996).

Kite et al. (1996) examined whether environmental auditors displayed significantly higher DIT *P* scores than did internal and external auditors. They posited that auditors with high DIT *P* scores who choose to leave public accounting might be attracted to job situations where they believe they can better serve the public interest, such as environmental auditing.

Kite et al found no significant differences between the DIT *P* scores for environmental auditors, internal auditors and previous samples of external auditors. When they further examined the DIT *P* scores of environmental auditors, Kite et al. discovered that persons requesting assignment to environmental audits displayed significantly higher levels of moral reasoning than did persons assigned by management. In citing several environmental disclosure studies, Kite et al reasoned that many companies wish to ignore environmental issues that require additional expenditure and public disclosure. Towards this end,

management may assign to environmental audits subordinates who are willing to put the interests of the firm ahead of the interests of society. Drawing upon the work of Ponemon (1992), Kite et al. suggested that firm socialization practices played a key role in the staffing of the environmental audit function.

Once again, the explanation that political ideology and not moral developmental differences is the critical factor in the selection/socialization process is highly plausible. Environmental protection, even when the cost may be detrimental to business interests, has been widely documented as a liberal cause. The higher observed DIT *P* scores for those persons requesting environmental audit assignments may result from liberal auditors being more likely to request such assignments. Conversely, the lower DIT *P* scores observed for internal auditors assigned to environmental audits may be the result of management assigning conservative auditors in expectation that they will be more likely to emphasize the interests of the business over those of the environment. Kite et al. (1996, p. 207) allude to this possibility by suggesting that management, while not being able to directly observe the moral reasoning levels of employees, can determine which internal auditors are most likely to protect company interests "through observing the employee's actions, political affiliations, and outside interests". Although they did not control for political ideology, Kite et al. suggest that it may be an important variable in the selection process.

DISCUSSION AND CONCLUSION

The results of this study, in conjunction with those of prior research (Emler et al., 1983; Sweeney & Fisher, 1998, 1999), provide strong evidence that the DIT *P* score, generated under the standard test instructions, confounds political ideology with moral reasoning development. As a result of the ideological bias in the DIT, the *P* scores of politically conservative and moderate test takers are likely understated in relation to their capacity for moral reasoning. Conversely, the *P* scores of liberal test takers are likely overstated.

In contrast to his earlier position, Rest et al. (1999a) recently admitted that the relationship between political ideology and DIT *P* scores is much stronger than initially believed. He conceded that political ideology explains as much as 40% of the variance in *P* scores, a result confirmed by Emler and Palmer-Canton (1998). Although acknowledging a relationship between the *P* score and political ideology, Rest and his colleagues (Rest et al., 1999a; Thoma et al., 1999) maintain that the *P* score does not reduce to a unitary measure of conservative or liberal preference and that the *P* score provides significant information beyond what may be obtained from knowing a person's political ideology.

This contention is not disputed. The multitude of studies demonstrating the developmental nature of moral reasoning, such as upward movement in DIT *P* scores with education and over time, support this view. Our concern is whether or not failure to address the significant overlap between political ideology and DIT *P* scores results in an unduly noisy measure of the moral reasoning construct, rendering the results of previous accounting studies incapable of being unambiguously interpreted. If political ideology is systematically influencing the test outcome, then failure to control for political ideology when employing DIT *P* scores in accounting studies would mask the true relationship between moral reasoning ability and examined behavior. While Rest and his colleagues have not openly conceded that political ideology should be controlled when using the DIT, the recently revised instrument, the DIT-2 (Rest et al., 1999b; Rest & Narvaez, 1998), now requires respondents to indicate their political ideology.

There are two important implications from this study for accounting ethics researchers. First, caution must be exercised in interpreting the results from previous empirical ethics studies that utilized the DIT. Results that had initially been attributed to a relationship between moral reasoning and important variables of interest may actually have been driven by political ideology. Perhaps in their zeal to adopt a widely used, objective measure, accounting researchers chose to ignore the existing evidence that the DIT was potentially a biased metric of moral reasoning ability (Emler et al., 1983). The results of this study should serve as a caveat to accounting researchers intending to import behavioral measures from outside disciplines.

Second, political ideology is likely a more important socialization variable in public accounting than previously recognized (Sweeney et al., 2001). Political ideology may effectively capture individual differences related to important attitudes, such as organizational and professional commitment, and to actions such as promotion, job assignment, performance appraisals, turnover and adherence to codes of conduct. Political ideology may also influence the self-selection processes of students considering accounting as a major and career.

A potential limitation of this study is the necessity of changing the DIT instructional set in order to examine its validity. Rest et al. (1999a) contend that studies manipulating DIT test instructions are incapable of revealing the impact of political ideology on *P* scores because the experimental condition elicits a response schema that is different from what occurs in normal test taking conditions. Rest had once been a proponent of faking studies (McGeorge, 1975; Hau, 1990; Bloom, 1977) as part of the validation process (Rest, 1979, 1993). Through this recent about face, Rest et al. (1999a) are attempting to create an

unassailable fortress around the DIT built on circular logic. By claiming that "manipulation of test instructions does not illuminate how people naturally or usually perform a task", Rest et al. (1997, p. 20) assume that the standard DIT instructions elicit a test-taker's best notions of justice and fairness (Rest 1979; 1986; 1993). Presented with evidence to the contrary, Rest and his colleagues (Rest et al., 1997, p. 20) dismiss each unflattering result by claiming that it ". . . is not evidence that participants usually take the DIT that way". But by not providing evidence that the standard instructions in fact elicit the normal or usual manner of reasoning, Rest et al. commit a circular folly. Understanding whether or not the instructions elicit the subject's most advanced moral schema free of potential systematic biases, such as political ideology, is critical to assessing the validity of the DIT.

NOTES

1. A consistent verdict of DIT research is that the moral reasoning abilities of accountants are less developed than those of college graduates and many other professional groups (Gaa, 1992).

2. There has been much criticism, rebuttal, and support for the philosophical underpinnings of Kohlberg's theory of cognitive moral development (see for example Kohlberg, 1981, Broughton, 1986; Weinreich-Haste, 1986; Schweder, 1982; Rest et al., 1999a). We do not wish to enter the debate. Our purpose is merely to assert that Kohlberg intended to link higher stages of moral thinking with political liberalism.

3. Contrary to Barnett et al. (1995), conservative and moderate subjects showed only small changes in their *A* scores. The "extremely liberal" manipulation was apparently less likely to draw subjects to anti-establishment items than the "left-wing radical" manipulation of Emler et al. (1983).

4. Similar results occurred when Sweeney and Fisher (1998) tested a DIT metric (the *N* score) proposed by Rest as an alternative to the *P* score.

ACKNOWLEDGMENTS

The authors would like to express their appreciation for the helpful comments from the editor, two anonymous reviewers, Finley Graves, David Donnelly, Stacy Kovar, participants at the *Central States Accounting Workshop*, and workshop participants at Washington State University, Kansas State University, and the University of Central Florida. Earlier versions of this paper were presented at the *1999 Critical Perspectives on Accounting Conference,* the *1999 Western Region of the American Accounting Association Annual Meeting,* and the *American Accounting Association Sixth Symposium on Ethics.*

REFERENCES

Barnett, R., Evens, J., & Rest, J. (1995). Faking moral judgment on the Defining Issues Test. *British Journal of Social Psychology, 34*(3), 267–278.

Bloom, R. (1977). Resistance to faking on the Defining Issues Test of moral development. Unpublished manuscript, College of William & Mary, Williamsburg, VA.

Broughton, J. M. (1986). The genesis of moral domination. In: S. Modgil & C. Modgil (Eds), *Lawrence Kohlberg: Consensus and Controversy* (pp. 363–385). Philadelphia, Pennsylvania: Falmer Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings.* Boston, MA: Houghton Mifflin Company.

Emler, N. P., Renwick, S., & Malone, B. (1983). The relationship between moral reasoning and political orientation. *Journal of Personality and Social Psychology, 45*(5), 1072–1080.

Emler, N., & Palmer-Canton, E. (1998). Politics, moral reasoning and the Defining Issues Test: A reply to Barnett et al. (1995). *British Journal of Social Psychology, 37*(4), 457–476.

Emler, N., & Stace, K. (1999). What does principled vs. conventional moral reasoning convey to others about the politics and psychology of the reasoner? *European Journal of Social Psychology, 29*(3), 455–468.

Eynon, G., Hill, N. T., & Stevens, K. T. (1997). Factors that influence the moral reasoning abilities of accountants: Implications for universities and the profession. *Journal of Business Ethics, 16*(9), 1297–1309.

Fisher, D. G., & Sweeney, J. T. (1998). The relationship between political attitudes and moral judgment: Examining the validity of the Defining Issues Test. *Journal of Business Ethics, 18*(6), 905–916.

Fishkin, J., Keniston, K., & MacKinnon, C. (1973). Moral reasoning and political ideology. *Journal of Personality and Social Psychology, 27*(1), 109–119.

Fontana, A. F., & Noel, N. (1973). Moral reasoning in the university. *Journal of Personality and Social Psychology, 27*(3), 419–429.

Gaa, J. C. (1992). The auditor's role: The philosophy and psychology of independence and objectivity. In: *Proceedings of the 1992 Deloitte & Touche/Unversity of Kansas Symposium on Auditing Problems* (pp. 7–43). University of Kansas.

Haan, N., Smith, M., & Block, J. (1968). Moral reasoning of young adults. *Journal of Personality and Social Psychology, 10*(3), 183–201.

Hau, K. T. (1990). Moral development and the ability to fake in a moral judgment test among Chinese adolescents. *Psychologia, 33*(2), 106–111.

Louwers, T. Ponemon, L., & Radtke, R. (1997). Examining accountants ethical behavior: A review and implications for future research. In: V. Arnold & S. Sutton (Eds), *Behavioral Accounting Research: Foundations and Frontiers* (pp. 188–221). Sarasota, FL: American Accounting Association.

Jeffrey, C., & Weatherholt, N. (1996). Ethical development, professional commitment, and rule observance attitudes: A study of CPAs and corporate accountants. *Behavioral Research in Accounting, 8*, 8–30.

Kite, D., Louwers, T. J., & Radtke, R. R. (1996). Ethics and environmental auditing: An investigation of environmental auditors' levels of moral reasoning. *Behavioral Research in Accounting, 8*(Supplement), 200–214.

Kohlberg, L. A. (1969). Stage and sequence: The cognitive developmental approach to socialization. In: D. A. Goslin (Ed.), *Handbook of Socialization Theory and Research* (pp. 347–480). Chicago, IL: Rand McNally.

Kohlberg, L. A. (1981). *Essays in Moral Development: The Philosophy of Moral Development,* (Vol. 1). New York: Harper & Row, Publishers.

Kohlberg, L. (1983). Moral development does not mean liberalism as destiny: A reply to Schweder. *Contemporary Psychology, 28*(1), 80–82.

Kohlberg, L. A., & Kramer, R. (1969). Continuities and discontinuities in childhood and adult moral development. *Human Development, 12*(1), 93–120.

Markoulis, D. (1989). Political involvement and socio-moral reasoning: Testing Emler's interpretation. *British Journal of Social Psychology, 28*(3), 203–212.

McGeorge, C. (1975). The susceptibility to faking of the Defining Issues Test of moral development. *Developmental Psychology, 11*(1), 108.

Nassi, A. J., Abramowitz, S. I., & Youmans, J. E. (1983). Moral development and politics a decade later: A replication and extension. *Journal of Personality and Social Psychology, 45*(5), 1127–1135.

Ponemon, L. A. (1992). Ethical reasoning and selection-socialization in accounting. *Accounting, Organizations and Society, 17*(3/4), 239–258.

Ponemon, L. A., & Gabhart, D. R. (1994). Ethical reasoning research in the accounting and auditing professions. In: J. R. Rest & D. Narvaez (Ed.), *Moral Development in the Professions: Psychology and Applied Ethics* (pp. 101–119). Hillsdale, New Jersey: Laurence Erlbaum Associates.

Rest, J. R. (1976). New approaches in the assessment of moral judgment. In: T. Lickona (Ed.), *Moral Development and Behavior* (pp. 198–220). New York: Holt, Rinehart & Winston.

Rest, J. R. (1979). *Development in judging moral issues. Minneapolis, Minnesota:* University of Minnesota Press.

Rest, J. R. (1986). *Moral development: Advances in research and theory.* New York: Prager Press.

Rest, J. R. (1993). *Guide for the Defining Issues Test,* version 1.3. Minneapolis, Minnesota: University of Minnesota.

Rest, J. R. (1994). Background: theory and research. In: J. R. Rest & D. Narvaez (Eds), *Moral Development in the Professions: Psychology and Applied Ethics* (pp. 1–26). New Jersey: Laurence Erlbaum Associates.

Rest, J., & Narvaez, D. (1998). *Guide for DIT-2,* version 2.3. Minneapolis, MN: University of Minnesota.

Rest, J., Narvaez, D., Bebeau, M. J., & Thoma, S. J. (1999a). *Postconventional moral thinking: A neo-Kohlbergian approach.* New Jersey: Lawrence Erlbaum Associates.

Rest, J. R., Narvaez, D., Thoma, S. J., & Bebeau, M. J. (1999b). DIT-2: Devising and testing a revised instrument of moral judgment. *Journal of Educational Psychology, 91*(4), 644–659.

Rest, J. R., Thoma, S. J., & Edwards, L. (1997). Designing and validating a measure of moral judgment: Stage preference and stage consistency approaches. *Journal of Educational Psychology, 89*(1), 5–28.

Schweder, R. (1982). Review of Lawrence Kohlberg's 'Essays in moral development', Vol. 1. *Contemporary Psychology, 27*(6), 421–424.

Sparks, P., & Durkin, K. (1987). Moral reasoning and political orientation: The context sensitivity of individual rights and democratic principles. *Journal of Personality and Social Psychology, 52*(5), 931–936.

Sweeney, J. T. (1995). The moral expertise of auditors: An exploratory analysis. In: L. Ponemon (Ed.), *Research in Accounting Ethics* (Vol. 1, pp. 213–234). Greenwich, CT: JAI Press.

Sweeney, J. T., & Fisher, D. G. (1998). An examination of the validity of a new measure of moral judgment. *Behavioral Research in Accounting, 10,* 138–158.

Sweeney, J. T., & Fisher, D. G. (1999). Politics, faking, and self-presentation: How valid is the *P* score of the Defining Issues Test as a measure of moral judgment? In: L. Ponemon (Ed.), *Research on Accounting Ethics* (Vol. 5, pp. 51–75). Greenwich, CT: JAI Press.

Sweeney, J. T., Quirin, J. J., & Fisher, D. G. (2001). Political ideology as an in-group prototype and socializing force in public accounting. Working paper: Washington State University, Pullman, WA.

Thoma, S. J., Narvaez, D., Rest, J., & Derryberry, P. (1999). Does moral judgment development reduce to political attitudes or verbal ability? Evidence using the Defining Issues Test. *Educational Psychology Review, 11*(3), 325–341.

Thorne, L., Massey, D. W., & Magnan, M. (2000). Insights into selection-socialization in the audit profession: An examination of the moral reasoning of public accountants in the United States and Canada. Working paper: York University, North York, Ontario.

Weinreich-Haste, H. (1986). Kohlberg's contribution to political psychology: A positive view. In: S. Modgil & C. Modgil (Eds), *Lawrence Kohlberg: Consensus and Controversy* (pp. 337–361). Philedelphia, PA: The Falmer Press.

THE EFFECT OF PRODUCT AND PROCESS COMPLEXITY ON PARTICIPATIVE LEADERSHIP STYLE

B. Douglas Clinton and Hossein Nouri

ABSTRACT

The purpose of this study is to examine the effects of process and product complexity on participative leadership style (i.e. perceptions of the primary decision maker in a resource allocation decision). Data from a descriptive judgment task provide evidence that product and process complexity operate as antecedents to choosing differential participative leadership decision styles. Analysis uses the Vroom Jago model (1988) to show that situations characterized by high (low) process complexity provide an increased (decreased) incentive to employ participative leadership styles. However, situations characterized by high (low) product complexity do not appear to significantly affect choices regarding participative leadership style.

INTRODUCTION

A review of budgetary participation research shows that most studies in this area have examined the relationship between budgetary participation and outcome variables such as performance (Brownell, 1981; Brownell & Dunk, 1991; Brownell & McInnes, 1986; Brownell & Merchant, 1990; Govindarajan, 1986;

Advances in Accounting Behavioral Research, Volume 5, pages 161–181.

ISBN: 0-7623-0953-9

Kren, 1992; Merchant, 1984; Mia, 1988, 1989; Nouri & Parker, 1998) and budgetary slack (Chow et al., 1988; Dunk, 1993; Govindarajan, 1986; Merchant, 1985; Nouri & Parker, 1996; Waller, 1988; Young, 1985). The results of these studies, in general, indicate that budgetary participation is an important element of the budget process. However, these studies make the assumption that budgetary participation exists and, then, attempt to relate it either directly or through moderating or mediating factors to outcome variables. Shields and Shields (1998, p. 50) argue that a potential reason for diverse results of the budgetary participation research could be due to the fact that "strong theoretical and empirical links between their assumed reason for why participative budgeting exists and their dependent variables" are absent from the extant research.

The purpose of this study is to examine the effects of specific antecedents (i.e. process and product complexity) on participative leadership style as perceived by the *primary* decision maker (superior) in a resource allocation decision. Several central issues suggest the importance of further examining the antecedent effects of participative leadership style: (1) prior studies concentrated on subordinate participation while superiors (i.e. leaders) *not* subordinates ultimately determine the type and amount of budgetary participation allowed in organizations; (2) there needs to be a fit between product and process complexity and the leadership style for optimal outcomes (e.g. Job performance); (3) since specific antecedents influence the perceived need for participative leadership styles and managerial choice regarding those styles, then considering these antecedents becomes important to properly interpreting consequences.

The Vroom Jago (VJ) model (1988), which classifies leadership styles as autocratic, consultative, and participative decision making, is used to examine the differential importance of process and product complexity antecedents. Based on a laboratory experiment, our findings show that superiors employ different leadership styles of participation in a resource allocation situation depending on the level of process complexity. Under high process complexity, superiors preferred group (Participative) decision making. Under low process complexity, superiors employed autocratic decision making. These findings appear to be supported by prior Vroom model studies, by studies examining the effects of perceived expertise, and by popular technological complexity definitions (e.g. Margerison & Glube, 1979; Clinton, 1996; Perrow, 1967).

The remainder of this study is organized as follows. The next section presents a literature review and includes formal hypotheses and theoretical support for hypothesized relationships. Next, the research method used in the study is described followed by the results section. Finally, conclusions, limitations, and a discussion of implications are presented.

LITERATURE REVIEW AND HYPOTHESES DEVELOPMENT

Vroom and Jago Model

Leadership style of budgetary participation can vary in degree (Locke & Schweiger, 1979; Swieringa & Moncur, 1975; Vroom & Jago, 1988). Table 1 presents different leadership styles in a budgetary participation context based on the Vroom and Jago (1988) leadership model.[1] The Vroom model is a normative contingency model that uses knowledge of situational factors such as decision quality and goal congruence to suggest one of five levels of subordinate participation in decision making. As Table 1 indicates, budgetary participation can vary from no participation (autocratic) to consultation to full participation (group consensus). The Vroom model suggests that the level of budgetary participation allowed would depend on situational factors coincident with a problem or decision faced by the leader (Vroom & Jago, 1988).

In accounting, Pasewark and Welker (1990) used the Vroom model (decision attributes, decision-making style, and decision rules) to investigate subordinate participation employed in budgeting decisions. They found support for the model in that high levels of participation increased the chance of successful budgetary decision making in instances where high participation was prescribed by the model. Clinton et al. (1996) showed that, consistent with the VJ model, the situational variable *task-specific superior/subordinate expertise* could behave as an important moderator of participation outcomes.

Table 1. Leadership Style in a Budgeting Context.

Leadership style	Who is involved	Who makes decision	Type of budget
AI. Autocratic	Leader	Leader	Assigned budgets (unassisted)
AII. Autocratic with information collection	Leader and others individually	Leader	Assigned budgets (individuals respond to specific questions)
CI. One-on-one consultation	Leader and others individually	Leader	Motivated budgets (individuals provide data and recommendations)
CII. Group consultation	Leader and others in group	Leader	Motivated budgets (group shares data and analyzes)
GII. Group consensus	Leader and others in group	Group	Full participative budgets (group shares data and analyzes)

Adapted from Vroom and Jago (1988).

Antecedents of Participative Leadership Style

Prior research shows that a variety of factors can contribute to the different levels and types of participation allowed in different contexts. Brownell and McInnes (1986), in explaining the inconsistent findings of their study, suggested that superiors might allow higher budgetary participation by those subordinates who perform well on the job. Beehr and Love (1983) suggested that there could be a feedback loop from performance to participation with a moderating effect of leadership style. Levine (1979) argued that budgetary participation might vary depending on organizational structure. Swieringa and Moncur (1975) noted that organizational variables might dictate what type of budgetary participation should be used.

Understanding why superiors use a particular style of participation is important because different leadership styles can affect outcome variables differently. Although not in a budgeting context, Cotton et al. (1988) found that direct participation as well as informal participation increase employee job performance while indirect participation through representatives had no effect on job performance. They further found that the impact of consultative participation on job performance was inconclusive.

The type of leadership style chosen can depend upon antecedent factors at virtually every level: organizational factors (e.g. technological complexity, organizational size), situational factors (e.g. information asymmetry, geographical dispersion), and individual factors (e.g. personality characteristics, and attitudes). These factors affect the type of leadership style used in budgetary participation (Briers & Hirst, 1990; Clinton, 1999).

A few prior studies have investigated antecedents of budgetary participation (e.g. Bruns & Waterhouse, 1975; Clinton, 1999; Clinton et al., 1996; Licata et al., 1986; Merchant, 1981, 1984; Mia, 1987; Shields & Young, 1993; Seiler & Bartlett, 1982). The findings of these studies, for example, show that the amount of participation depends on how much control supervisors perceived was needed (Bruns and Waterhouse 1975), whether internal managers allowed higher subordinate budgetary participation than external managers (Licata et al., 1986), and whether there was a negative relationship between need for independence and extent of participation (Seiler & Bartlett, 1982). The evidence from prior empirical studies provides support for the significance of antecedent variables. Fellner and Sulzer-Azaeoff (1985) examined the effect of assigned and participative goal setting in a manufacturing environment. They found that participative goal setting was not as effective as assigned goals in augmenting the safety performance of employees. These findings suggest that depending on the antecedent(s) of budgetary participation, an autocratic leadership style may

be more useful than a participative style. Certainly in other cases methods that are more participative appear appropriate (Cotton et al., 1988).

Technological Complexity

Brownell (1982) was possibly the first accounting study to formally include technology in a model depicting antecedents of participation in budgeting. In the broader participative decision making literature, two Vroom model studies (Margerison & Glube, 1979; Paul & Ebadi, 1989) held technological complexity constant while conducting validation studies of both the descriptive and normative abilities of the model. This emphasizes the importance of considering technological complexity in the choice of leadership styles.

There are many definitions and taxonomies of technology, but Perrow's (1967, p. 195) definition of technology as "the actions that an individual performs upon an object, with or without the aid of tools or mechanical devices, in order to make some change in that object" is perhaps the most widely accepted. Under this definition, technology is the tools, techniques, or knowledge being used to transform inputs into outputs (Daft & Lengel, 1986).

Studies in the accounting literature have operationalized technological complexity as *task uncertainty* (Brownell & Hirst, 1986, Brownell & Dunk, 1991). Moreover, studies in organization theory that have not specifically operationalized technology as task uncertainty have used definitions that emphasize the effects of technology on individuals (Perrow, 1967; Thompson, 1967; Fry, 1982; Withey et al., 1983; Daft & Lengel, 1986). These studies use disaggregated technology constructs.

Disaggregated technology constructs (e.g. analyzability and variety) were originally theoretically developed and explored by Perrow (1967) and further examined in studies such as Daft and Lengel (1986) in their discussion of the divergent effects of uncertainty and ambiguity as produced by technological complexity. This study uses product and process complexity as subconstructs of technology to examine the effects of technology on leadership styles.

Product and Process Complexity

Product complexity refers to product problems or activities that cannot be standardized or predicted in advance. For complex products, component parts are of a wide variety, differing in function, material, and specification (products are custom made). Under product complexity, changes in the parts are often made due to continual technological advances (Brownell & Merchant, 1990).

Process complexity refers to situations where specialized tools are needed to diagnose problems. For complex processes, the employees are highly skilled and constantly interact with each other, and their jobs are difficult and involve a large variety of important decisions. A complex process requires a substantial amount of judgment and creativity. Table 2 presents organizational examples of technological complexity as we dichotomize these two continua into low and high product complexity and low and high process complexity.

To understand the expected differential effects for cells shown in Table 2, first consider the work environments of a warranty service center in comparison to a pharmaceuticals manufacturer. Tasks of the service center involve trouble-shooting and determining a course of action by listening to convoluted and often imprecise tales of problems as related by customers. In this sense, the service center process is difficult to analyze and can often involve a considerable amount of judgment and creativity. In a similar fashion, diagnostic approaches may not be considered repetitive or routine since every customer and situation is different. However, the product serviced may be reasonably non-complex. That is, all units produced could have the same materials and specifications and be designed to perform the same function. Component parts could be interchangeable and not subject to frequent modification in design or configuration. This would provide an example of a complex environment involving a high degree of process complexity, but yet a low degree of product complexity (cell 2 in Table 2).

In comparison, an examination of the operations of the pharmaceuticals manufacturer could reveal a routine manufacturing process (in the Perrow sense). That is, drugs may be produced in a clear and orderly sequential process. Automated equipment could render worker activity simple and repetitive and

Table 2. Organizational Examples of Technological Complexity.

		Process Complexity	
		Low	**High**
Product Complexity	**Low**	1 Assembly	2 Warranty Service
	High	3 Component Manufacturing	4 Research & Development

require little or no worker judgment or creativity. In fact, the precision involved with accurately producing drugs may *require* the use of automation and that process inputs and outputs be intricately specifiable. However, the drugs being manufactured may vary widely in materials and specifications. Properties of the materials may be comparatively unique and not well understood to the operators. Medical breakthroughs and advances may require frequent changes in product characteristics. In this example, the operating environment would be characterized as routine (low complexity) on the process technology dimension, yet involve extreme product complexity (cell 3 in Table 2). Characteristics of the other cells in Table 2 (i.e. cells 1 and 4) can be generalized from combinations of these descriptions.

Two forces appear to be driving the effect of process complexity on leadership style. First, participatory leadership style is desirable as a means of gathering information and resolving ambiguity (Daft & Lengel, 1986). Second, the element of relatively high expertise is likely to be present, or perceived to be present, coincident with the presence of complex processes. To use the subordinates' expertise and to overcome ambiguity, superiors may use participatory (group) leadership style.

The role of participatory leadership style in this sense, consistent with Daft and Lengel (1986), can be to increase the quantity of information as extracted from subordinates and/or to increase the richness of the information transfer between the budget decision maker and the subordinate. Consistent with the prior discussion, the presence of high process complexity requires subordinates with the requisite expertise to handle the complexity. These subordinates will thus have an enhanced ability to meaningfully participate in budgeting over those in low process complexity contexts that typically lack expertise. The related hypothesis, stated in alternative form, is:

> *H1*: Perceptions of high process complexity will cause decision makers to prefer a more participative leadership style in a resource allocation situation than when process complexity is perceived to be relatively low.

For low product complexity, the optimal input/output relation is either known or can be learned easily, thus eliminating the need for a participatory leadership style (Brownell & Merchant, 1990). By contrast, complex products are of a wide variety, differing in function, material, and specification. There would be "difficulty in unambiguously specifying the proper set and order of input" (Brownell & Merchant, 1990). Participatory leadership style should assist in resolving these ambiguities (Brownell & Merchant, 1990). This leads to the second hypothesis of this study, stated in alternative form:

H2: Perceptions of high product complexity will cause decision makers to prefer a more participative leadership style in a resource allocation situation than when product complexity is perceived to be relatively low.

METHOD

Subjects and Observations

Forty-three students in one cost accounting course participated in an in-class laboratory experiment as partial fulfillment of a class requirement. Average age of the subjects was 24, and 38% of the subjects were male. The class was comprised of juniors and seniors. Using four trials for each subject resulted in 172 usable observations on which to base inferences.

Experimental Design and Procedures

The two factors of *process complexity* and *product complexity* were manipulated in the study. Each factor had two levels: (1) low and (2) high. All manipulations were contained within the scenario descriptions detailing the degree of product and process complexity within each trial.

The subjects were told that they were participating in a judgment exercise that required them to assume the role of a production supervisor. All other instructions were in writing and included in a folder that was identical for all subjects. Subjects were required to perform the task in one sitting and were instructed to not discuss the exercise with others or allow anyone else to influence their responses.

After reading a page detailing a general problem situation, each subject completed four trials consisting of four different cases providing additional information about the general problem. To complete each trial, subjects answered several questions, four of which provided manipulation checks relevant to each trial. After completing the four trials, subjects completed a post-exercise questionnaire containing 11 questions pertaining to demographic data and the overall exercise.

Experimental Task

Subjects were presented with a problem situation that described a resource allocation decision within a budgetary context (see Appendix A). Subjects were not required to actually make the allocation decision, but were required to

indicate how they would go about making the decision in terms of leadership style. That is, they indicated what type of leadership style they would use in each scenario. The Vroom-Yetton-Jago scale[2] was used to measure the leadership styles.[3] They also answered questions designed to provide manipulation checks for product and process complexity (see Appendix B).

The four cases were presented in counterbalanced order and described four different divisions in a computer manufacturing company.[4] Descriptions for each division consisted of two paragraphs, one describing the product environment and one describing the process environment. Four plants were selected which fulfilled the four unique factor-level combinations of high and low product and process complexity. The case descriptions were based on the work of Perrow (1967), Daft and Lengel (1986), Withey et al. (1983), Mintzberg (1979), and Harvey (1968) and are presented in Appendix C.

RESULTS

Manipulation Checks

To evaluate the success of the manipulations of product and process complexity, four manipulation check questions were administered after each trial as repeated-measures questions.[5] Pairs of questions for product and process complexity were summed to form the specific level of perceived product and process complexity. A t-test ($t = 23.10$, $p < 0.001$) indicated that subjects in the high process complexity trials perceived more process complexity ($\xi = 8.08$, $s = 1.668$) than in the low process complexity trials ($\xi = 3.07$, $s = 1.125$). The correlation between the two questions specific to process complexity was 0.872 ($p < 0.001$).

In addition, a t-test ($t = 5.77$, $p < 0.001$) indicated that subjects in the high product complexity trials perceived more product complexity ($\xi = 6.86$, $s = 2.635$) than those in the low product complexity trials ($\xi = 4.74$, $s = 1.625$). The correlation between the two questions specific to product complexity was 0.800 ($p < 0.001$). Accordingly, the manipulations of both product complexity and process complexity appeared to have been successful.[6]

Hypothesis Tests

Although the VJ model reveals five levels of participation, the creators show how the levels are *not* equidistant but rather seem to break into three pieces. They label these three leadership styles as autocratic (AI and AII), consultative (CI), and group (CII and GII) methods (Vroom & Jago, 1988, p. 96). Extensive testing

Table 3. Frequency Table Presenting Cases by Leadership Style Selected.

Leadership Style	Frequency	Percent	Cumulative Percent
Autocratic	76	44.2	44.2
Consultative	22	12.8	57.0
Group (Participatory)	74	43.0	100.0

by VJ of the five levels shows that scaling attributed provides considerable distance between the three general level-types. Their results ascribed consistent level scoring on a zero to ten scale of AI = 0, AII = 1, CI = 5, CII = 8, and GII = 10. This scaling reveals large differences between level CI and the polar extremes on either end, but this same degree of separation is not present for AI to AII or CII to GII. This suggests that for the purpose of this study, the use of three levels (autocratic, consultative, and group leadership styles) is adequate.

The VJ discussion (Vroom & Jago, 1988, p. 37) regarding the three levels is consistent with the hypotheses presented in this study.

> Autocratic leadership methods require different skills than more participative methods. To employ AI effectively, a leader must be intelligent enough to make a high-quality decision. It is also helpful if the leader is 'charismatic' or, at the very least, skilled in the art of persuasion. Autocratic methods or decision processes require skills both in decision making and in inspiring others.

This quotation suggests the importance (once again) of expertise and the subordinate's perceptions regarding expertise. They go on to say . . . "Group methods such as GII and CII require a very different set of skills, namely the skills of a facilitator or discussion leader" (Vroom & Jago, 1988, p. 37).

A frequency table presenting cases by style selected is presented in Table 3. It shows that participants in the experiment mainly selected autocratic or group (participatory) leadership style.

To test the study hypotheses and to accommodate the three-level categorical dependent variable (leadership style), multinomial logistic regression, a special case of the general log-linear model, was used (Demaris, 1992; Knoke & Bruke, 1980; Menard, 1995). Let p_G, p_C, and p_A be the probabilities that a participatory, consultative, and autocratic leadership style is chosen, respectively ($p_G + p_C + p_A = 1$). The odds of a case dropping into a given category, such as autocratic style, are provided by the probability ratios p_G / p_A and p_C / p_A. The multinomial logistic regression assumes that the dependent variable, odd-logs of one choice relative to another, can be expressed as a linear function of a number of variables related to the dependent variable. The models to be estimated in this study are:

$$\log (p_G / p_A) = \alpha_1 + \beta_1 \text{ PROCESS} + \delta_1 \text{ PRODUCT}$$
$$\log (p_C / p_A) = \alpha_2 + \beta_2 \text{ PROCESS} + \delta_2 \text{ PRODUCT}$$
$$\log (p_G / p_C) = \alpha_3 + \beta_3 \text{ PROCESS} + \delta_3 \text{ PRODUCT}$$

Table 4 presents the regression of the three levels of leadership style on the process and product complexity.[7]

The findings in Table 4 show that the models' chi-square statistics are significant ($p < 0.001$). The pseudo R^2, as estimated by the Nagelkerke (1991) formula, indicate that about 30% of the variation in the leadership style is explained by the multinomial logistic regression model. The results further show that process complexity is significant in all three models supporting *H1*. Since product complexity coefficients were not significant, *H2* is not supported.

Table 4. Results of Multinomial Logistic Regression.

Variables	Coefficient	STD Error	Chi-Square	*P*	Exp (Coeffi.)
log (p_G / p_A) Participatory vs. autocratic style					
INTERCEPT	−1.359	0.345	15.515	0.000	
PROCESS	2.471	0.393	39.607	0.000	11.839
PRODUCT	0.350	0.383	0.836	0.361	1.419
log (p_C / p_A) Consultative vs. autocratic style					
INTERCEPT	−1.558	0.401	15.515	0.000	
PROCESS	1.453	0.514	7.979	0.005	4.276
PRODUCT	−0.455	0.514	0.785	0.376	0.634
log (p_G / p_C) Participatory vs. consultative style					
INTERCEPT	0.200	0.458	0.190	0.663	
PROCESS	1.018	0.512	3.958	0.047	2.768
PRODUCT	0.806	0.507	2.526	0.112	2.238

Model Chi-Square = 50.392 with 4 df, $p = 0.000$

Percent correctly classified:

Participatory leadership style	76.3%
Consultative leadership style	0.0%
Autocratic leadership style	78.4%
Total	67.4%

Sample size = 172

PROCESS = Process complexity.
PRODUCT = Product complexity.

The results in Table 4 indicate that the ratio of the odds of choosing group (participatory) leadership style in contrast to autocratic leadership style is 11.8 times higher for high process complexity than for low process complexity. The ratio of the odds of choosing consultative vs. autocratic leadership style is about 4.3 times higher when process complexity is high than when process complexity is low. Finally, the ratio of the odds of choosing group (participatory) vs. consultative leadership style is about 2.8 times higher for high process complexity than for low process complexity.

Table 4 also shows that the model correctly classified 76.3% of group (participatory) leadership style, none of consultative leadership style, and 78.4% of autocratic leadership style. Overall, about 67.4% of the leadership style choices are correctly assigned. These results in general indicate that, as process complexity increases, leaders in resource allocation situations tend to employ leadership styles that are more participatory.

CONCLUSIONS, LIMITATIONS, AND DISCUSSION

Consistent with recent calls to examine antecedents of budgetary participation (Shields & Shields, 1998; Shields & Young, 1993), this study examines two antecedents of budgetary participation, namely process and product complexity. An experimental design was used to observe the effect of the constructs of process and product complexity on perceived leadership style in a hypothetical resource allocation situation. We used the Vroom-Jago model (1988) to show that process complexity is associated with selection of participative (group) leadership styles. Product complexity, however, did *not* appear to significantly affect choices regarding participative (group) leadership style.

Product complexity was not evidenced to be as important in affecting leadership style as *process* complexity. These findings indicate that superiors may not perceive that participatory (group) decision making is necessary under high product complexity. Brownell and Merchant (1990) showed that budgetary participation interacts with product complexity to affect job performance such that for high product complexity, budget participation led to a more positive performance. The practical implication in light of our findings is that superiors may not employ a participatory leadership style under high product complexity where in fact such a leadership style could be beneficial.

Under high *process* complexity, the superior needs to gather enough information to deal with the non-routine nature of the work. Since subordinates working under process complexity are assumed to be experienced or otherwise possess requisite capacity to perform complex processes, the leader may use a more participatory budgetary style to take advantage of that experience and

expertise. This suggests that when process complexity is high the leader would be more likely to choose a style consistent with group consensus (Cells 2 and 4 in Table 2).

In interpreting the findings of this study, one should also consider its limitations. Because of the experimental nature of the study, the results may not be generalizable to all settings. This is particularly important when we consider that in field settings, other factors may also come into play when one decides what types of leadership style to use. For example, Seiler and Bartlett (1982) showed that need for independence decreases the extent of participation. Since in a high process complexity situation leaders prefer a participatory style, need for independence may hamper the use of such leadership style. In addition, subjects in this study were students who may not represent budget managers. Future studies could replicate this study in a more natural setting. Finally, this study uses a within-subjects experimental design. A within-subjects design is more susceptible to experimental demands and guessing of hypotheses by participants.

Despite its limitations, this study shows that different leadership styles may be employed in resource allocation situations depending on the level of process complexity present. This finding may further help to explain why some prior studies did not find a relation between budget participation and job performance. In fact, if budgetary participation is assumed to exist regardless of process complexity, no significant association between budget participation and job performance should exist. This is because the autocratic leadership style that is appropriate for low process complexity necessitates the use of an assigned budget but does not affect the subordinate's performance. The same is true for high process complexity where participatory budgeting is employed and, again, the subordinate's performance is not affected. Since both assigned and participatory budgets lead to similar performance, no relation between budget participation and job performance would exist.

In summary, results of the study reveal that process complexity can increase the choice of participation in budgeting tasks. Moreover, evidence is provided to cast doubt on the presumption that product complexity affects the choice of participatory leadership style.

NOTES

1. The *Vroom and Jago model* or merely the *Vroom model* is used here and elsewhere in this paper to refer to either or both the Vroom and Yetton model (1973) or the revised Vroom and Jago model (1988). Where the distinction is important, they are denoted by their more specific model names.

2. The *Vroom-Yetton-Jago* scale refers to that used by either the Vroom-Yetton model (1973) or the revised Vroom-Jago model (1988), and it is presented as question 1 in Appendix B.

3. The participative decision-making literature reveals that the Vroom scale has been extensively tested for both reliability and validity and has been used in more than 60 participative decision making studies spanning greater than 20 years. Therefore, reliability testing for the Vroom scale was regarded as unnecessary in the current study.

4. Pilot testing revealed that subject responses were not significantly associated with the order of presentation of the cases, case paragraphs, or case questions.

5. Wording of the questions was adapted from items loading high on the refined scale of technological complexity produced by Withey et al. (1983) as modified for the technology concepts of complexity and variety per Harvey (1968) and sophistication per Mintzberg (1979). These questions are presented as items 2 through 5 in Appendix B. The scale developed by Withey, et al. was factor analyzed and showed convergent and divergent validity as well as construct validity as verified by the independent pooled responses of individuals with experience in both teaching and research using Perrow's model.

6. One may argue that process complexity is confounded with employee type. That is, participants in the experiment view employees in high process complexity situation as highly skilled, etc. and as employees who attach more importance to being able to participate in important decisions of the plant. Or one may argue that subjects may have perceived that to secure commitment of highly skilled employees in contrast to low skilled employees, participative leadership style should be used. While as with all experiments these and other scenarios are possible, the manipulation check clearly indicates that subjects have also distinguished between high and low process complexity.

7. To test whether subjects were influenced by the general case description rather than the complexity issues we added a dummy variable for the specific case into the models. The dummy variables were not significant, nor were there significant changes on the results reported in this study.

8. Source: Vroom, V. H., and A. G. Jago. 1988. *The New Leadership: Managing Participation in Organizations*. Englewood Cliffs, NJ: Prentice Hall. Reader note: This citation was not present on the original questionnaire.

9. Italicized case names shown here did not appear on the original forms.

ACKNOWLEDGMENTS

We are grateful for the helpful suggestions and comments of several individuals on earlier drafts of this paper. Accordingly, we would like to thank Thomas W. Hall, John M. Hassell, Sally A. Webber, an anonymous doctoral student at the University of Texas at Arlington, and participants of the AAA annual meeting and the national meeting of the Institute for Operations Research and Management Sciences.

REFERENCES

Beehr, T. A., & Love, K. G. (1983). A meta-model of the effects of goal characteristics, feedback, and role characteristics in human organizations. *Human Relations*, *36*(2), 151–166.

Briers, M., & Hirst, M. (1990). The role of budgetary information in performance evaluation. *Accounting, Organizations and Society*, *15*(4), 373–398.

Brownell, P. (1981). Participation in budgeting, locus of control and organizational effectiveness. *The Accounting Review*, *56*(October), 844–860.

Brownell, P. (1982). Participation in the budgeting process: When it works and when it doesn't. *The Journal of Accounting Literature*, *1*, 124–150.

Brownell, P., & Dunk, A. S. (1991). Task uncertainty and its interaction with budgetary participation and budget emphasis: Some methodological issues and empirical investigation. *Accounting, Organizations and Society*, *16*(8), 693–703.

Brownell, P., & Hirst, M. (1986). Reliance on accounting information, budgetary participation, and task uncertainty: Tests of a three-way interaction. *Journal of Accounting Research*, (Autumn), 241–249.

Brownell, P., & McInnes, M. (1986). Budgetary participation, motivation, and managerial performance. *The Accounting Review*, *61*(October), 587–600.

Brownell, P., & Merchant, K. A. (1990). The budgetary and performance influences of product standardization and manufacturing process automation. *Journal of Accounting Research*, (Autumn), 388–397.

Bruns, W. J., Jr., & Waterhouse, J. H. (1975). Budgetary control and organization structure. *Journal of Accounting Research*, *13*(Autumn), 177–203.

Chow, C. W., Cooper, J. C., & Waller, W. S. (1988). Participative budgeting: Effects of a truth-inducing pay scheme and information asymmetry on slack and performance. *The Accounting Review*, (January), 111–122.

Clinton, B. D. (1999). Antecedents of budgetary participation: The effects of organizational, situational, and individual factors. *Advances in Management Accounting*, *8*, 45–70.

Clinton, B. D., Hall, T. W., Hunton, J. E., & Pierce, B. J. (1996). Perceptions of supervisor/subordinate task-specific expertise as a moderator of participative budgeting outcomes. *Advances in Accounting*, *14*, 85–106.

Cotton, J. L., Vollrath, D. A., Froggatt, K. L., Lengnick-Hall, M. L., & Jennings, K. R. (1988). Employee participation: Diverse forms and different outcomes. *Academy of Management Review*, *13*(1), 8–22.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science*, 554–571.

Demaris, A. (1992). Logit modeling: Practical applications. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-086. Newbury Park, CA: Sage.

Dunk, A. S. (1993). The effect of budget emphasis and information asymmetry on the relation between budgetary participation and slack. *The Accounting Review*, (April), 400–410.

Fellner, D. J., & Sulzer-Azaroff, B. (1985). Occupational safety: Assessing the impact of adding assigned or participative goal-setting. *Journal of Organizational Behavior Management*, *7*(1/2), 3–24.

Fry, L. W. (1982). Technology-structure research: Three critical issues. *Academy of Management Journal*, *25*, 532–552.

Govindarajan, V. (1986). Impact of participation in the budgetary process on managerial attitudes and performance: Universalistic and contingency perspectives. *Decision Sciences, 17*, 496–516.

Harvey, E. (1968). Technology and the structure of organizations. *American Sociological Review, 33*, 247–259.

Knoke, D., & Bruke, P. J. (1980). Log-linear models. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-020. Beverly Hills, CA: Sage.

Kren, L. (1992). Budgetary participation and managerial performance: The impact of information and environmental volatility. *The Accounting Review, 67*(July), 511–526.

Levine, M. (1979). Participative budgeting – reviewing the literature. *Business*, (March-April), 49–52.

Licata, M. P., Strawser, R. H., & Welker, R. B. (1986). A note on participation in budgeting and locus of control. *The Accounting Review*, (January), 112–117.

Locke, E. A., & Schweiger, D. M. (1979). Participation in decision-making: One more look. In: B. M. Staw (Ed.), *Research in Organizational Behavior, 1* (pp. 265–339). Greenwich, CT: JAI Press.

Margerison, C., & Glube, R. (1979). Leadership decision-making: An empirical test of the Vroom and Yetton model. *The Journal of Management Studies*, (February), 45–55.

Menard, S. (1995). Applied logistic regression analysis. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-106. Thousand Oaks, CA: Sage.

Merchant, K. A. (1981). The Design of the corporate budgeting system: influences on managerial behavior and performance. *The Accounting Review, 56*(October), 813–829.

Merchant, K. A. (1984). Influences on departmental budgeting: An empirical examination of a contingency model. *Accounting, Organizations and Society, 9*(3/4), 291–307.

Merchant, K. A. (1985). Budgeting and the propensity to create budgetary slack. *Accounting, Organizations and Society, 10*(2), 201–210.

Mia, L. (1987). Participation in budgetary decision making, task difficulty, locus of control, and employee behavior: An empirical study. *Decision Sciences, 18*, 547–561.

Mia, L. (1988). Managerial attitude, motivation and the effectiveness of budget participation. *Accounting, Organizations and Society, 13*, 465–476.

Mia, L. (1989). The impact of participation in budgeting and job difficulty on managerial performance and work motivation: A research note. *Accounting, Organizations and Society, 14*(4), 347–357.

Mintzberg, H. (1979). *The Structuring of organizations: A synthesis of the research.* Englewood Cliffs, NJ: Prentice Hall.

Nagelkerke, N. J. D. (1991). A note on general definition of the coefficient of determination. *Biometrika, 78*, 691–692.

Nouri, H., & Parker, R. J. (1996). The effect of organizational commitment on the relation between budgetary participation and budgetary slack. *Behavioral Research in Accounting, 8*, 74–90.

Nouri, H., & Parker, R. J. (1988). The relationship between budget participation and job performance: The roles of budget adequacy and organizational commitment. *Accounting, Organizations and Society, 23*(5/6), 467–483.

Pasewark, W. R., & Welker, R. B. (1990). A Vroom-Yetton evaluation of subordinate participation in budgetary decision making. *Journal of Management Accounting Research, 2*, 113–126.

Paul, R. J., & Ebadi, Y. M. (1989). Leadership decision making in a service organization: A field test of the Vroom-Yetton model. *Journal of Occupational Psychology, 62*, 201–211.

Perrow, C. (1967). A Framework for the comparative analysis of organizations. *American Sociological Review, 32*, 194–208.

Seiler, R. E., & Bartlett, R. W. (1982). Personality variables as predictors of budget system characteristics. *Accounting, Organizations and Society, 7*, 381–403.

Shields, J. F., & Shields, M. D. (1998). Antecedents of participative budgeting. *Accounting, Organizations and Society, 23*(1), 49–76.

Shields, M., & Young, S. M. (1993). Antecedents and consequences of participative budgeting: Evidence on the effects of asymmetrical information. *Journal of Management Accounting Research, 5*, 265–280.

Swieringa, R. J., & Moncur, R. H. (1975). *Some effects of participative budgeting on managerial behavior*. New York, NY: National Association of Accountants.

Thompson, J. D. (1967). *Organizations in action*. New York, NY: McGraw-Hill.

Vroom, V. H., & Jago, A. G. (1988). *The new leadership: Managing participation in organizations*. Englewood Cliffs, NJ: Prentice Hall.

Waller, W. S. (1988). Slack in participative budgeting: The joint effect of a truth-inducing pay scheme and risk preferences. *Accounting, Organizations and Society, 13*(1), 87–98.

Withey, M., Daft, R. L., & Cooper, W. H. (1983). Measures of Perrow's work unit technology: An empirical assessment and a new scale. *Academy of Management Journal, 26*(1), 45–63.

Young, S. M. (1985). Participative budgeting: The effects of risk aversion and asymmetric information on budgetary slack. *Journal of Accounting Research*, (Autumn), 829–842.

APPENDIX A

General Problem Situation

In each of the cases that follow, you will play the role of a plant manager in a computer manufacturing division of a large company. The division manager has allocated $100,000 to each plant this year to spend on tools, equipment, and the bonuses of twenty employees that are at each plant. Basically, these twenty employees are all at the same level (in terms of merit) at each plant. The division manager has told you that as plant manager, you are required to decide how the $100,000 is to be allocated among the tools, equipment, and employees. The objective, according to the division manager, is for you to allocate the money to achieve the greatest overall effectiveness for your plant. Therefore, the technical quality of the decision is very important. There is a general feeling of competitiveness between the other plant managers and yourself. Also, there is an expectation that the division manager will be assessing the performance of all plant managers to see which one will produce the best results. For this reason it is extremely important to carefully decide how the $100,000 will be distributed.

You have decided that the only viable investment alternative for tools and equipment for each plant is a particular package costing $95,000. The items in the package cannot be purchased separately, so the only decision is to purchase

the package or not. The new tools and equipment would definitely be expected to enhance the performance of the plant, and the plant could purchase these tools and equipment leaving $5,000 to be paid in bonuses to the twenty employees ($250 each). On the other hand, financial performance of the company has been sluggish, and the employees have been working under conditions of a wage freeze for the past three years. With this in mind, it would be nice to give the employees the full bonus of $100,000 ($5,000 each), but that would mean foregoing the purchase of the tools and equipment. Either way, you could provide the employees with a bonus, but you are concerned that if you merely give them the $250 each, they might be insulted at receiving such a small amount. Under normal conditions you believe that the purchase of the tooling and equipment would produce the best results for the firm. The employees have been asking for the equipment for a long time. However, the employees have been asking for salary adjustments and bonuses for a long time also. You are concerned, given these circumstances, that the decision (either way) could seriously affect employee morale thus affecting the performance of the plant. Therefore, employee commitment to the decision is very important and must be carefully considered. Although the two alternatives available to you are clear, the uncertainty regarding which action will produce the greatest overall effectiveness for your plant makes the problem extremely unstructured and difficult to resolve.

The information above will remain consistent from case to case. However, you will be asked to assume some additional information about the nature of the situation that is different from case to case. Please consider this information when answering the eleven questions for each case. These questions address alternative approaches available to you in making your decision involving varying approaches to participation and communication with the employees you supervise.

APPENDIX B

Repeated Measures Case Questions

(1) Indicate the method that you feel would be most appropriate for this case.

(a) You choose to solve the problem by making the decision yourself using the information available to you at the time.
(b) You obtain any necessary information from subordinates, then make the decision yourself. You may or may not tell subordinates the purpose of your questions or give information about the problem or decision on which you are working.

(c) You share the problem with the relevant subordinates individually, getting their ideas and suggestions without bringing them together as a group. Then *you* make the decision. This decision may or may not reflect your subordinates' influence.
(d) You share the problem with your subordinates in a group meeting. In this meeting you obtain their ideas and suggestions. Then *you* make the decision, which may or may not reflect your subordinates' influence.
(e) You share the problem with your subordinates as a group. Together you attempt to reach agreement (consensus) on a decision. You do not try to "press" them to adopt "your" decision, and you are willing to accept and implement any decision that has the support of the entire group.[8]

(2) In terms of job difficulty, sophistication, and task variety or **process** complexity, how would you describe the relative level of technological complexity that the employees experience at this plant?
(3) In terms of complexity, sophistication, and variety of the **product(s)**, how would you describe the relative level of technological complexity that the employees experience at this plant?
(4) In terms of the routine or non-routine nature of employee activity regarding the plant **process**, how would you describe the relative level of technological complexity that characterizes this plant's environment?
(5) Regarding the routine or sophisticated nature of employee activity regarding the plant's **product(s)**, how would you describe the relative level of technological complexity that characterizes this plant's environment?

APPENDIX C

Cases

Case One – Component Manufacturing Plant[9]

This plant manufactures component parts for computers. Components are produced in a clear and orderly sequential process. The equipment used makes worker activity simple and repetitive and allows no deviation in procedures. The employees are low-skilled operators and rarely interact. Their jobs are considered simple and involve repetitive, routine decisions. The assembly process requires no judgment, craftsmanship, or creativity on the part of the workers.

Component parts produced are of a wide variety, differing in function, materials, and specifications. Many times they are customized for a particular application. Often the materials are specialized, or are otherwise unique, and

have properties that are not well-understood by the employees. Changes in the parts are often made due to continual technological advances.

Case Two – Research & Development
This plant is responsible for new product development. The process cannot be described by a step-by-step orderly sequence. Specialized tools are often needed to build the production prototypes. The employees are highly-skilled engineers, scientists, and technicians and constantly interact. Their jobs are considered difficult and involve a large variety of important decisions. The development process requires a substantial amount of judgment, craftsmanship, and creativity involving high sophistication and complex intricacy.

Production prototypes are often sophisticated and can vary considerably from one to another. Many times they are customized for a particular target market. They often involve many parts that are created from special materials. Changes in the direction of developmental efforts are made constantly as relevant technological advances are made.

Case Three – Assembly Plant
This plant assembles a single model of finished computers. Basic personal computers are assembled using a clear and orderly sequential process. The equipment used makes worker activity simple and repetitive and allows no deviation in procedures. The employees are low-skilled operators and rarely interact. Their jobs are considered simple and involve repetitive, routine decisions. The assembly process requires no judgment, craftsmanship, or creativity on the part of the workers.

The finished computers produced all have the same materials and specifications and are designed to perform alike. The computers are general purpose units and will likely be used by consumers. Parts used are made of routine or common materials and are assembled with ordinary fasteners. The assembly of the units is rarely affected by technological changes since new parts are simply inserted into a common case.

Case Four – Warranty Service Plant
This plant performs warranty repair service. The process cannot be described by step-by-step orderly sequence. Specialized tools are often needed to diagnose problems and perform repairs. The employees are highly-skilled diagnosticians and technicians and constantly interact. Their jobs are considered difficult and involve a large variety of important decisions. The trouble shooting and repair process requires a substantial amount of judgment, craftsmanship, and creativity involving high sophistication and complex intricacy.

The computers requiring repair generally all have the same materials and specifications and are designed to perform similar functions. Computers eligible for warranty service at this plant are general purpose units sold to, and used by consumers. Parts used for these units are routine or standardized to be interchangeable for repair. Repair parts are not typically affected by technological changes since old parts are stocked to use in existing units.

THE EFFECTS OF BUDGET EMPHASIS, PARTICIPATION AND ORGANIZATIONAL COMMITMENT ON JOB SATISFACTION: EVIDENCE FROM THE FINANCIAL SERVICES SECTOR

Chong M. Lau and Jason Chong

ABSTRACT

Research on the impact of performance evaluative style on managers' behavior suggests that organizational interest is best served if a high budget emphasis evaluative style is used in a high participatory environment, whilst a low budget emphasis evaluative style is used in a low participatory environment. However, research results are not always consistent. This may be due to the omission of the influence of organizational commitment on managers' behavior. Highly committed managers are likely to strive for organizational goals and interests whereas lowly committed managers are likely to strive for personal goals and interests. These conflicting attitudes are likely to affect the relationships among budget emphasis, budgetary participation and managers' behavior. Highly committed managers, striving for organizational goals, may react favorably to the compatible combinations

Advances in Accounting Behavioral Research, Volume 5, pages 183–211.

ISBN: 0-7623-0953-9

of high (low) budget emphasis and high (low) budgetary participation. In contrast, lowly committed managers, striving for personal goals, may prefer other combinations of budget emphasis and budgetary participation. Based on a sample of 112 financial services sector managers, these expectations are supported by the results of this study.

INTRODUCTION

Much research has been devoted to the area of supervisory evaluative styles (budget emphasis) since the seminal studies by Hopwood (1972) and Otley (1978). The continuing stream of studies in this and the related areas of budgetary participation indicates that these areas are regarded by many researchers as important and contemporary issues in management accounting research (Kren & Liao, 1988; Briers & Hirst, 1990; Lindsay & Ehrenberg, 1993; Otley & Fakiolas, 2000). For instance, Brownell and Dunk (1991, p. 703) observed that "the continuing stream of research devoted to this issue constitutes . . . the only organized critical mass of empirical work in management accounting." Lindsay and Ehrenberg (1993, p. 223) similarly suggest that these research areas constitute "one of the relatively few areas in management accounting research where there has been any sequence of repeated studies." Continuous research in these areas is not surprising, considering the importance of budgets in the planning, control and performance evaluation systems of most contemporary organizations. Whilst there has been much interest in non-financial performance indicators since Kaplan's (1983) study, they are not intended to replace financial performance indicators. Rather, they are intended to supplement and use jointly with financial performance indicators. For instance, financial performance indicators remain an important category of performance indicators in the Balanced Scorecard developed by Kaplan and Norton (1996). Horngren et al. (2000, p. 468) suggest that "the balanced scorecard places strong emphasis on financial objectives and measures . . . (and) emphasizes non-financial measures as a part of a program to achieve future financial performance." Moreover, the use of non-financial indicators, as performance evaluation criteria, may not be radically different from the superiors' evaluative styles that were examined in prior studies, which generally categorized evaluative styles into high budget emphasis and low budget emphasis. High budget emphasis generally refers to evaluative styles that place heavy reliance on financial and accounting measures, whilst low budget emphasis refers to evaluative styles that rely on non-financial and non-accounting criteria (Hopwood, 1972; Otley, 1978). The use of non-financial performance indicators is therefore closely related to this research area. Consequently,

research evidence from studies in the area of supervisory evaluative styles is still relevant to contemporary organizations. Additionally, planning, control and performance evaluation are always needed, regardless of whether financial or non-financial performance evaluation criteria are used. Targets (budgets) for non-financial performance indicators are needed and need to be developed in a similar manner as financial targets. Control systems, which compare actual performance with planned performance, whether expressed in financial or non-financial terms, are just as essential today as in the past. In their review of studies on supervisory evaluative style, Otley and Fakiolas (2000, p. 509) concluded as follows:

> . . . the design of studies of the impact of accounting techniques in performance measurement and management function should ensure that they are able to pick up . . . changing emphasis . . . *But performance measurement and the use of performance measures in performance evaluation are still key management issue.* The use of the Balanced Scorecard technique is proving very popular . . . . Nevertheless, the methods used in studies on RAPM and evaluative style appear to be generalizable into this arena with only relatively minor adaptation. *The door is open for researchers to build upon this foundation and make a central contribution to the management accounting literature.*

Hence, there is strong justification to continue to extend the research in the area of supervisory evaluative style. Based on the results of prior studies, some important general patterns have emerged. In particular, prior studies have found that supervisory evaluative style (budget emphasis) and budgetary participation interact to affect managers' behavior and performance. This means that beneficial behavioral outcomes (e.g. improved managerial performance or job satisfaction) may be associated with certain combinations of budget emphasis and budgetary participation. Brownell (1982) suggested that beneficial behavioral outcomes are found when a *high* budget emphasis evaluative style is complemented by *high* budgetary participation, and when a *low* budget emphasis evaluative style is complemented by *low* budgetary participation.

Other studies have extended on these findings by investigating the effects of other moderating variables such as task uncertainty (Brownell & Hirst, 1986), task difficulty (Brownell & Dunk, 1991; Lau et al., 1995) and national culture (Harrison, 1992); and on other dependent variables such as job-related tension (Brownell & Hirst, 1986), and job satisfaction (Harrison, 1992). Overall, whilst these results are relatively consistent with respect to the *high* budget emphasis and *high* participation combination, they are equivocal with respect to the *low* budget emphasis and *low* participation combination. In particular, whilst both Brownell and Dunk (1991) and Lau et al. (1995) found support for the *high* budget emphasis-*high* participation combination, their results with respect to the *low* budget emphasis-*low* participation combination were in the opposite

direction. Hence, whilst Brownell and Dunk (1991) found the *low* budget emphasis-*low* participation combination to be associated with favorable behavioral outcome, Lau et al. (1995) found this combination to be associated with unfavorable behavioral outcome. Hence, there is a need to resolve these inconsistencies in research results.

Our study suggests that these inconsistencies in results could be explained by omitted variables, one of which is organizational commitment. It proposes that managers, who are highly committed to their organizations, are likely to react differently from those who are lowly committed to their organizations. Highly committed managers are likely to place great importance on their organizations' interests and work towards the attainment of these interests. Consequently, the expectation is that the theory pertaining to the beneficial consequences of a compatible combination of high (low) budget emphasis and high (low) participation is likely to hold for such managers. In contrast, for managers, who are lowly committed to their organization, it is likely that they will place their self-interests ahead of their organizations' interests. Consequently, the theory pertaining to the beneficial consequences of a compatible combination of high (low) budget emphasis and high (low) participation may not necessarily hold for these managers. The omission of organizational commitment in both Brownell and Dunk (1991) and Lau et al. (1995) may have caused the discrepancies in the results of these two studies pertaining to the low budget emphasis-low participation combination. To date, the moderating effects of organizational commitment on the relationship among budget emphasis, budgetary participation and managers' behavior have remained largely untested (Nouri, 1994; Nouri & Parker, 1996).

This study therefore investigates the three-way interaction between budget emphasis, budgetary participation and organizational commitment affecting managers' job satisfaction in the financial services sector. Figure 1 presents the model used in this study. The financial services sector is studied here because even though it is one of the major sectors of most economies, it has generally been neglected by management accounting researchers (Pope & Otley, 1996; Lau & Tan, 1998). Job satisfaction is studied here because it is one of the major behavioral outcomes examined in management accounting research in general, and in the area of supervisory evaluative style in particular. Harrison (1992, p. 8) regards job satisfaction as an "individually important outcome in (its) own right and also been seen as leading to organizationally important outcomes including absenteeism and turnover, motivation, job involvement and performance."

In the next section, relevant studies are examined to develop a theoretical basis for the hypotheses to be tested. Subsequent sections, respectively, describe the method, results and the consequential implications for theory and practice.

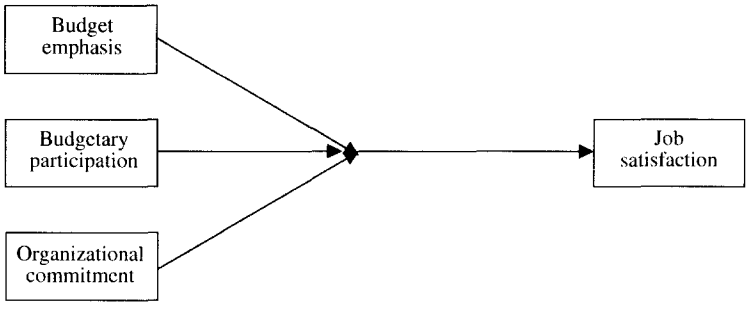


Fig. 1. Three-Way Interaction Between Budget Emphasis, Budgetary Participation and Organizational Commitment Affecting Job Satisfaction.

HYPOTHESIS DEVELOPMENT

Based on the principle of operant conditioning and balance theory, Brownell (1982) theorized that a match between high (low) budget emphasis and high (low) budgetary participation is crucial for beneficial behavioral outcomes to occur. He argued as follows:

> Operant conditioning suggests that the orientation of the subordinate toward the evaluative style of his superior is dependent on the level of budgetary participation in the planning phase. A favorable orientation toward those evaluative styles which place a high emphasis on budget results in the control phase will occur only under conditions of high participation. Similarly, a favorable orientation toward a low budget emphasis style will occur only under conditions of low budgetary participation. An unfavorable orientation thus results either from high budget emphasis/low participation or low budget emphasis/high participation. The predicted consequences of differing orientations toward the superior's' evaluative style are based on balance theory. Specifically, both performance and reported levels of job satisfaction should be lower with inappropriate than with appropriate degrees of emphasis placed by the superior on budget results (Brownell, 1982, p. 14).

Hence, the organization's interest is best served if managers, who are evaluated with a high budget emphasis evaluative style, are allowed high budgetary participation, whilst managers, evaluated with a low budget emphasis evaluative style, are allowed only low budgetary participation.

Brownell's theory regarding the beneficial effects of compatible combinations of budget emphasis and participation was re-examined and refined in a number of subsequent studies (e.g. Brownell & Hirst, 1986; Brownell & Dunk, 1991; Lau et al., 1995). In particular, both Brownell and Dunk (1991) and Lau et al. (1995) predicted and found support for Brownell's (1982) finding of the aforementioned significant two-way interaction between budget emphasis and

budgetary participation affecting managers' behavior. Brownell and Dunk found that: (i) the *high* budget emphasis-*high* participation combination; and (ii) the *low* budget emphasis-*low* participation combination were *both* associated with improved behavioral outcome. Lau et al. found that the *high* budget emphasis-*high* participation combination was associated with beneficial outcome. These results were consistent with Brownell and Dunk. However, unlike Brownell and Dunk, Lau et al. found that the *low* budget emphasis-*low* participation was *not* associated with beneficial consequences. Instead, they found this combination to be associated with the worst behavioral outcome.

In summary, a compatible match of *high* budget emphasis and *high* participation being associated with favorable behavioral consequences was supported by all three studies (i.e. Brownell, 1982; Brownell & Dunk, 1991; Lau et al., 1995). However, a compatible match of *low* budget emphasis and *low* participation being associated with favorable consequences was supported by Brownell (1982) and Brownell and Dunk (1991), but *not* by Lau et al. (1995). These inconsistencies in results could be attributed to omitted variables. Brownell's (1982) theory and all the aforementioned subsequent studies have overlooked the importance of organizational commitment on managers' behavior and attitudes.

The Role of Organizational Commitment

Organizational commitment is an important variable in management and organizational behavior studies (Sommer et al., 1996). However, within the context of budget related studies, organizational commitment has not been examined extensively (Nouri, 1994; Nouri & Parker, 1996).

There are several different conceptualizations of organizational commitment in the literature. These different conceptualizations arise because of the existence of at least two dimensions, namely affective or attitudinal commitment and continuance or calculative commitment, that are captured by the organizational commitment concept (Mathieu & Zajac, 1990; Randall, 1990; Meyer et al., 1990; Meyer & Allen, 1991; Nouri & Parker, 1996). Mowday et al. (1979, p. 225) regard affective or attitudinal commitment as "a state in which an individual identifies with a particular organization and its goals and wishes to maintain membership in order to facilitate these goals." In contrast, Meyer and Allen (1991, p. 64) describe continuance commitment as "the continuation of an action (e.g. remaining with an organization) resulting from a recognition of the costs associated with its termination." Meyer et al. (1990, p. 710) conclude that "employees with a strong affective commitment remain with the organization because they want to, whereas those with strong continuance commitment remain because they need to."

The affective or attitudinal concept of organizational commitment is employed in this study. Randall (1990, p. 369) considered that "attitudinal/moral conceptualization of OC [organizational commitment] has a stronger relationship with work outcomes than a calculative conceptualization of OC." Porter et al. (1974), Mowday et al. (1979), Angle and Perry (1981), and Mowday et al. (1982) all associate affective commitment with three characteristics: (i) a strong belief in and acceptance of organization's goals; (ii) a willingness to exert considerable effort on behalf of the organization; and (iii) a definite desire to maintain organizational membership. Mathieu and Zajac (1990) also found that affective commitment was related to job satisfaction. Using the affective/attitudinal concept is consistent with those of prior budget related studies (Nouri, 1994; Nouri & Parker, 1996, 1998).

Prior literature suggests that individuals with different levels of organizational commitment may have different utility functions. As noted by Lincoln and Kallenberg (1990, p. 22):

> The committed employee's involvement in the organization . . . extends beyond the satisfaction of merely personal interest in employment, income, and intrinsically rewarding work. The employee becomes conscious of the needs of the organization and how his or her actions contribute to the fulfilment of those needs. To identify with the organization, then, implies that the worker is willing to expend effort for the sake of the company . . .

This is consistent with Nouri (1994) and Nouri and Parker (1996, 1998) who found that managers with different levels of organizational commitment demonstrate different attitudes towards their organizations. They showed that managers with high organizational commitment view organizational goals as important, and therefore the pursuit of these goals and objectives are of utmost importance. In contrast, managers with low organizational commitment are primarily concerned with pursuing self-interests and place little or no value in achieving organizational goals and objectives. These differences in objectives may have important effects on the managers' attitudes and behavior. First, a number of researchers (eg. Steer, 1977; Bateman & Strasser, 1984; DeCotiis & Summers, 1987; Poznanski & Bline, 1997) have argued that managers, who are highly committed to their organizations, may experience higher levels of job satisfaction than managers who are lowly committed to their organizations. Second, and importantly for this study, managers with different levels of organizational commitment are likely to react differently to different combinations of budget emphasis and budgetary participation. This means that organizational commitment is not only likely to be positively associated with job satisfaction, but is also likely to influence the two-way interaction between budget emphasis and participation affecting job satisfaction. This is discussed further below.

Interactive Effects of Organizational Commitment

Prior research results on the impact of performance evaluative style on managers' behavior indicate that organizational interest is best served if a *high* budget emphasis evaluative style is used in a *high* participatory environment, whilst a *low* budget emphasis evaluative style is used in a *low* participatory environment (Brownell, 1982; Brownell & Dunk, 1991; Lau et al., 1995). Since managers with high organizational commitment are primarily concerned with pursuing organizational goals and interests, they are likely to accept these combinations of budget emphasis and budgetary participation because the acceptance of these combinations is likely to be in the best interest of their organizations. They may accept these combinations even if these combinations are not necessarily in their best *personal* interests because these managers may regard their personal performance and rewards as secondary to those of their organizations. *Consequently, these prior research results are likely to hold for the highly committed managers.*

In contrast, since lowly committed managers are likely to be primarily concerned with pursuing self-interests and place little or no value in achieving organizational goals, these prior research results on budget emphasis and budgetary participation may not hold because they may not be in the best self-interest of the managers. Since the pursuit of personal rewards and remunerations are likely to be of paramount importance to lowly committed managers in their pursuit of self-interest, they are likely to view their superiors' evaluations of their performance seriously because these performance evaluations are likely to be closely linked to their organizations' reward systems. Consequently, they may be very sensitive to the evaluative styles used by their superiors to evaluate their performance and are likely to prefer an evaluative style that best serves their self-interest of achieving favorable performance evaluations.

Lowly committed managers are likely to prefer a high budget emphasis evaluative style to a low budget emphasis evaluative style. With a low budget emphasis evaluative style, managers are likely to be evaluated by multiple non-accounting criteria, such as concern with quality, ability to get along with superiors and peers and ability to handle the work force (Hopwood, 1972; Otley, 1978). Ross (1994, p. 630) suggested that managers are likely to be suspicious and apprehensive of a *low* budget emphasis evaluative style because "the criteria associated with the non-accounting performance evaluation style are somewhat subjective and therefore, may well be ambiguous and difficult to measure" and are subject to "a superior's biases and idiosyncrasies". He further suggested that managers evaluated by a low budget emphasis style "may always have a

nagging doubts as to the reasonableness of an evaluation based on subjective criteria. . ..The managers may, therefore, feel a lack of control over their ability to affect their evaluation." In contrast, he considered evaluative styles based on accounting performance measures to be much more acceptable to managers because accounting based criteria are much more "objective" and "verifiable" than non-accounting criteria. Empirically, Hopwood (1972, p. 173) found that trust was positively associated with the two criteria of "meeting the budget" and "concern with costs', the two items he used to measure accounting based criteria. He suggested that "this might reflect the fact that one purpose of a budget is to clearly set out the objectives for a cost center. While this certainly cannot be done with perfect accuracy, it is possible to carefully and cautiously use the budget for this purpose and thereby add an important element of structure and clarity to the job environment." He further suggested that:

> A non-accounting evaluation, in particular, might be made on the basis of rather vague criteria: attitudes, the way the cost center head handles his men, and effort. Whilst such criteria are important, they are surrounded by a great deal of uncertainty. It is difficult to clearly specify what constitutes good and bad performance, and a supervisor might find it difficult to determine when improvement occurs. In these circumstances, the budgetary systems offer one definite advantage. It attempts to express the unit's objectives in a precise manner. (Hopwood, 1972, p. 174).

Even though the recent interest in non-financial performance indicators (Kaplan, 1983; Kaplan & Norton, 1996) has led to the development of quantifiable non-financial performance criteria, these criteria are still likely to be more subjective and ambiguous than accounting-based performance measures. Compared with accounting criteria, non-financial criteria are generally much harder to quantify. For instance, in their discussion on the Balanced Scorecard, Horngren et al. (2000, pp. 469–470) recommend the inclusion of what they regard as "subjective" measures such as customer satisfaction rating. However, they caution that "when using *subjective* measures, though, management must be careful to trade off the benefits of the richer information of these measures provide *against the imprecision and potential for manipulation*" (italics added). In addition, with non-accounting criteria, managers are usually required to satisfy more than one criterion (Kaplan & Norton, 1996; Horngren et al., 2000), which therefore involves arbitrary assignment of weights to the different criteria. Hence, *on balance*, there is likely to be a higher degree of subjectivity and ambiguity associated with a low budget emphasis evaluative style than with a high budget emphasis evaluative style.

The above therefore suggests that lowly committed managers, who may be much more sensitive to performance evaluation than highly committed managers, are likely to prefer a *high* budget emphasis evaluative style to a *low*

budget emphasis evaluative style. This means that, whilst a mismatched combination of *high* budget emphasis-low participation was considered by prior studies (Brownell, 1982; Brownell & Dunk, 1991) to be not in the organizations' interests, lowly committed managers may still prefer this combination to a *low* budget emphasis-low participation combination because they may be less suspicious of a high budget emphasis evaluative style than a low budget emphasis evaluative style.

In summary, prior studies (Brownell, 1982; Brownell & Dunk, 1991) suggest that managers prefer *compatible* combinations of: (i) *high* budget emphasis-*high* participation; and (ii) *low* budget emphasis-*low* participation to *other* combinations. This study suggests that managers, who are *highly committed* to their organizations, are likely to accept these combinations because they are in their organizations' best interests. However, because *lowly committed* managers may prefer a *high* budget emphasis evaluative style to a *low* budget emphasis evaluative style, they are likely to prefer combinations with *high* budget emphasis to combinations with *low* budget emphasis. Hence, whilst they may accept a *high* budget emphasis-high participation combination, they are likely to reject a *low* budget emphasis-low participation combination. Instead, they are likely to prefer a *high* budget emphasis-low participation combination to a *low* budget emphasis-low participation combination. *This means that, with respect to the **low** budget emphasis-**low** participation combination, whilst **highly** committed managers are likely to react **favorably** to this combination, **lowly** committed managers are likely to react **unfavorably** to this combination.* This suggests a three-way interaction between budget emphasis, budgetary participation and organizational commitment affecting job satisfaction. The following hypothesis is therefore tested:

> *H1*: There is a significant three-way interaction between budget emphasis, budgetary participation and organizational commitment affecting job satisfaction.

In order to examine the nature of this three-way interaction, the following discussion dichotomizes budgetary participation into two situations, namely, high and low budgetary participation situations.

High Budgetary Participation Situations

Since budgetary participation facilitates communication between the managers and their superiors, leading to a lower level of information asymmetry, the working environment is likely to be more open, transparent and trusting.

Consequently, there is likely to be a greater agreement by the managers with the evaluative styles used by their superiors to evaluate their performance. According to Otley (1978), it is the extent of the managers' agreement rather than any particular evaluative style that will affect the managers' behavior and attitudes. Since there is agreement with the evaluative styles used, whether low or high budget emphasis, the extent of budget emphasis may not have any significant effect on the managers' job satisfaction in a high budgetary participation environment.

Additionally, highly committed managers are likely to value their participation privileges as a tool to achieve organizational goals (Nouri & Parker, 1996). Because performance evaluations and personal rewards are secondary to this group of managers, they are likely to view budgetary participation as an opportunity to learn about their organizational goals and how achievements of budget targets can assist in the achievements of organizational goals. Even if budget emphasis is low, these managers may still find participation rewarding because it enables them to communicate with their superiors and learn about their organizations' affairs. Consequently, they are likely to experience high levels of job satisfaction with such participation in their organizations' affairs.

In contrast, because managers with low organizational commitment are interested in pursuing self-interests in the form of favorable job performance evaluations and rewards, they are likely to derive their satisfaction from what their jobs have to offer and not from their jobs themselves. Hence, whilst their superiors may intend participation to be used as a tool to increase the managers' motivation and morale towards their jobs (Cherrington & Cherrington, 1973; Collins, 1978; Merchant, 1981), these managers, whose only goals are in favorable job evaluations and rewards, are not likely to appreciate the benefits of such subordinate-superior interactions incorporated in their participation privileges. Instead, prior studies have found that these managers view participation as a chance to satisfy their personal interests and an opportunity to incorporate slack into their budgets (Nouri & Parker, 1996), making them easier to attain (Lowe & Shaw, 1968). Such dysfunctional activities are likely to reduce the extent of their job satisfaction.

In summary, lowly committed managers, who focus primarily on self-interests in the form of favorable evaluations and rewards from their jobs, are unlikely to find job satisfaction through participation in their organizational affairs. In contrast, highly committed managers, whose interests are in the achievement of organizational goals and objectives, are likely to make use of budgetary participation as a tool in helping them to satisfy their job interests. *It is therefore possible to conclude that in high budgetary participation situations, the managers' job satisfaction is likely to be positively associated with*

the extent of the managers' organizational commitment (Steer, 1977; Bateman & Strasser, 1984; DeCotiis & Summers, 1987; Poznanski & Bline, 1997). This leads to the following hypothesis:

H2: In high budgetary participation situations, organizational commitment is positively associated with job satisfaction.

Low Budgetary Participation Situations

Low organizational commitment

As discussed earlier, in situations where participation in setting budgets and communication with their superiors are low, *lowly committed* managers may be suspicious of the evaluative styles used by their superiors. With no or few opportunities to participate and clarify the evaluative criteria used by their superiors to evaluate their performance, a budget target may be easier to understand than such non-accounting targets as cooperation with colleagues, getting along with superior, concern with quality, ability to handle work force and attention towards the company (Hopwood, 1972; Ross, 1994). Consequently in low participatory environment, these managers are likely to prefer a high budget emphasis evaluative style and may experience a higher level of job satisfaction if they are evaluated with a high budget emphasis than if they are evaluated with a low budget emphasis. The following hypothesis is therefore tested:

H3: In low budgetary participation situations, *high* budget emphasis is associated with higher job satisfaction than low budget emphasis when organizational commitment is *low*.

High organizational commitment

In contrast, managers with high organizational commitment are likely to demonstrate a different behavior. Since these managers view the pursuit of organizational goals to be of outmost importance, they are likely to be less concerned than lowly committed managers with the evaluative styles that are used to evaluate their performance. Hence, it would be easier for them to accept the evaluative style used by their superiors to evaluate their performance. Since prior studies (Brownell, 1982; Brownell & Dunk, 1991) have suggested that the *low* budget emphasis evaluative style is compatible with *low* budgetary participation and in the organizations' best interest, it is likely that highly committed managers will find a low budget emphasis evaluative style to be more satisfying than a high budget evaluative style when budgetary participation

is low. These managers may view such a compatible combination to be in the best interests of their organizations. For instance, a manager in an organization, which emphasizes customer services, may not want to concentrate on the achievement of budget targets, but may want to provide the best possible services to all customers. Hence, Brownell's (1982) suggestion that managers will experience higher job satisfaction if a low budget emphasis evaluative style is used in low budgetary participation situations may be applicable to highly committed managers. Accordingly, the following hypothesis is tested:

H4: In low budgetary participation situations, *low* budget emphasis is associated with higher job satisfaction than high budget emphasis when organizational commitment is *high*.

METHOD

This study used a mailed questionnaire survey as data collection method. Questionnaires were sent to 200 functional heads from 61 financial services institutions, including insurance companies, finance houses and different categories of banks selected randomly from *Kompass Australia* 1998. In order to provide for some control over the size of organizations, only financial institutions with more than 100 employees were included. The names of the functional managers from the organizations were obtained by telephone calls to the organizations so that the questionnaires could be directed to them personally. A maximum of five managers' names were obtained from each organization so as to provide a cross-sectional representation of managers from the industry. Some organizations provided less than five names. In addition to the functional heads, a total of 31 branch bank managers were also included in the sample of 200 respondents. The branches included into the sample were randomly selected from the available customer service center branches available from the White Pages 1997/1998.

A copy of the questionnaire together with a self-addressed prepaid envelope and a covering letter, which assured the respondents of their confidentiality, were mailed to each respondent in August 1998. Respondents who did not reply were sent a follow-up letter three weeks later. A total of 115 responses were received. Three responses were not useable because two respondents indicated that budgets were not used in their organizations and one questionnaire was incomplete. This resulted in a total of 112 useable questionnaires, yielding a final response rate of 56%.

The managers who responded to the questionnaires had a mean age of 43.0 years. They had been in their current positions for an average of 3.7 years with an average of 54.8 number of employees working in their areas of responsibility. In addition, they had an average experience of 11.2 years in their respective areas of responsibility. With regards to their respective academic backgrounds, about 60% of the managers had a university degree or professional qualification or both.

Oppenheim's (1992) non-response bias test was employed in this study. T-tests were applied to the scores of all the variables of the early and late response groups. The results indicated that the scores of the independent and dependent variables between the two groups were all not significantly different. These results indicate that there are no significant differences between the early and late groups of responses.

Measurement Instruments

Budget emphasis

Budget emphasis was measured with Hopwood's (1972) modified eight-item, seven-point Likert-scaled version. It had been widely applied in prior research to measure superior's evaluative styles (e.g. Otley, 1978; Brownell, 1982, 1985; Brownell & Hirst, 1986; Brownell & Dunk, 1991; Harrison, 1992; Lau et al., 1995).

The first modification of the categorization came from Brownell (1982) when he collapsed the four evaluative styles into two classes, namely, high budget emphasis (incorporating Budget-conscious and Budget-profit) and low budget emphasis (incorporating Profit-conscious and Non-accounting). In order to overcome the problem of respondents failing to complete rank ordering properly faced by Brownell (1982), later researchers (e.g. Brownell, 1985; Brownell & Dunk, 1991) used the Likert-type rating scale to measure each of the items in the instrument. Hopwood (1972) had found the ranking method to be consistent with the rating method.

Consistent with prior studies (e.g. Brownell, 1985; Brownell & Dunk, 1991; Lau et al., 1995), the score for budget emphasis was obtained by adding the scores of the two accounting criteria, "My ability to meet budgeted targets in the short run" and "My long run concern with costs and/or revenues". This approach operationalized budget emphasis on a continuum from accounting to non-accounting style of evaluation. This was done following suggestions by Otley (1978) and Govindarajan (1984) that performance evaluative style is better

conceptualized as a continuous rather than a categorical variable. Brownell (1985, p. 505) stressed that "the legitimacy of this procedure is dependent on a significant correlation between the responses to the two items". In this study, the two criteria correlated at 0.508 ($p < 0.01$).

Organizational commitment

Organizational commitment was measured using the nine-item seven-point Likert-scaled instrument short-form version of Organizational Commitment Questionnaire (OCQ) developed by Mowday et al. (1979). Prior researchers have suggested that OCQ has good psychometric properties (e.g. Mowday et al., 1979, Angle & Perry, 1981) and that it is also the "most widely used measure of affective commitment to date" (Meyer et al., 1989, p. 152). Prior budget related studies found it to have acceptable level of internal consistency with Cronbach alpha, reported at 0.87 by Nouri (1994) and 0.86 Nouri and Parker (1996, 1998). Studies in other fields also reported acceptable levels of reliability and validity. For instance, Blau (1987) reported a Cronbach alpha of 0.87.

The Cronbach alpha for this instrument in this study is 0.92. The results of factor analysis for organizational commitment show that all items loaded on one factor with high factor loadings (> 0.5) except for item 3 which has a factor loading of 0.448, which is only marginally less than the benchmark of 0.5. Eigenvalue is 5.933 and the total variance explained is 65.93%.

Budgetary participation

Budgetary participation was measured using Milani's (1975) six-item, seven-point Likert-scaled instrument. An additive scale was used where the budgetary participation score was obtained by adding the scores of all six items. This instrument has been used extensively and proven to be internally reliable with high Cronbach (1951) alpha in prior research (eg. Brownell & Hirst, 1986; Brownell & Dunk, 1991; Harrison, 1992; Lau et al., 1995; Nouri & Parker, 1996, 1998). The Cronbach alpha obtained in this study is 0.89. The factor analysis indicates that all the items loaded satisfactorily into one factor with factor loadings of higher than 0.5. The eigenvalue is 3.895 and the total variance explained is 64.91%.

Job satisfaction

Job satisfaction was measured using the 20-item short-form of the Minnesota Satisfaction Questionnaire (MSQ) developed by Weiss et al. (1967). This instrument was also used by Harrison (1992, 1993), who suggested that "the

MSQ is the most comprehensive of the facet-specific measures of job satisfaction, comprising 20 job facets" (Harrison, 1993, p. 329). The short form instead of the long-form (100-item instrument) was used to keep the overall questioning of the subjects within a reasonable time frame, as the subjects (subordinate managers) were very busy people.

There are two main advantages of MSQ over other job satisfaction measurements (e.g. Job Descriptive Index (JDI) developed by Smith et al., 1969). Firstly, in analyzing different measures of job satisfaction, Dunham et al. (1977) found MSQ provided the highest level of convergent validity and outperformed other measures in tests of discriminant validity. Secondly, Scarpello and Campbell (1983) declared that MSQ has had more success than the other scales that they had investigated, in the prediction of overall job satisfaction from satisfaction with the facets of the job. In addition, Weiss et al. (1967) found the MSQ to be have high levels of reliability and validity when tested with an entire sample of 27 norm groups and a specific group of managers. This short version of the MSQ has not only been consistently used in budget related studies (e.g. Brownell, 1982; Frucot & Shearon, 1991; Harrison, 1992, 1993) but also in research from other disciplines (e.g. Butler, 1983). Within the budget-related literature, researchers also reported high Cronbach alphas associated with the short form MSQ. For instance, Harrison (1992) reported Cronbach alpha of 0.93.

Since MSQ measures the degree of managers' satisfaction with 20 different job facets, factor analysis was not appropriate for this variable and hence was not used. The Cronbach alpha in this study is 0.90, which is relatively high, indicating that the MSQ instrument is internally reliable and consistent. Descriptive statistics of both the independent and dependent variables are presented in Table 1.

Table 1. Descriptive Statistics of the Independent and Dependent Variables ($n = 112$).

Variables	Mean	Standard deviation	Possible range Min.	Possible range Max.	Actual range Min.	Actual range Max.	Cronbach alpha
Budget emphasis	11.95	2.03	2	14	2	14	n/a
Budget participation	30.57	8.36	6	42	8	42	0.89
Organizational commitment	48.83	10.21	9	63	13	63	0.92
Job satisfaction	72.54	11.17	20	100	33	93	0.90

RESULTS AND DISCUSSION

Hypothesis H1

Hypothesis H1 states that there is a significant three-way interaction between budget emphasis, budgetary participation and organizational commitment affecting job satisfaction. The following hierarchical regression model Eq. (1) was used to test hypothesis H1.

$$Yi = b0 + b1Bi + b2Pi + b3Oi + b4BiPi + b5BiOi + b6PiOi + b7BiPiOi + ei \quad (1)$$

where:

Yi = Job satisfaction.
Bi = Budget emphasis.
Pi = Budgetary participation.
Oi = Organizational commitment.
ei = Error term.

Table 2 presents the results of the three-way interaction between budget emphasis, budgetary participation and organizational commitment affecting job satisfaction. The results indicate that the coefficient b_7 of the three-way interaction is significant and positively related to job satisfaction (est = 0.016; $p < 0.013$). The coefficient of determination (R^2) is 40.5%. These results provide support for hypothesis H1.

Table 2. Results of Regression of Job Satisfaction on Budget Emphasis, Budgetary Participation and Organizational Commitment ($n = 112$).

	Equation (1)			
Variable	Coeff	Est	*t*-value	*p*
Constant	b_0	−286.680	−2.617	0.005
Budget emphasis (BE)	b_1	26.514	2.924	0.002
Budgetary participation (PA)	b_2	10.594	2.509	0.007
Organizational commitment (OC)	b_3	6.770	2.995	0.002
BE × PA	b_4	−0.823	−2.358	0.010
BE × OC	b_5	−0.513	−2.740	0.004
PA × OC	b_6	−0.196	−2.330	0.011
BE × PA × OC	b_7	0.016	2.266	0.013
R^2			0.405	
Adjusted R^2			0.365	
F value			10.098	
$p \leq$			0.001	

Hypotheses H2, H3 and H4

Recall that hypothesis H2 is related to the *high* participatory situations and states that organizational commitment is positively associated with job satisfaction in high budgetary participation situations. This means that in high budgetary participation situations, organizational commitment is expected to have a significant positive main effect on job satisfaction.

Hypothesis H3 and H4 are related to the low participatory situations. Hypothesis H3 states that in low budgetary participation situations, *high* budget emphasis is associated with higher job satisfaction than low budget emphasis *when organizational commitment is low.* In contrast, hypothesis H4 states that in low budgetary participation situations, *low* budget emphasis is associated with higher job satisfaction than high budget emphasis *when organizational commitment is high.* These two hypotheses suggest that in low budgetary participation situations, a significant two-way interaction between organizational commitment and budget emphasis is expected.

In order to enable the aforementioned three hypotheses to be tested, budgetary participation was dichotomized at its mean and the main effects and two-way interaction effects were respectively tested using the following models:

$$Yi = b0 + b1Bi + b2Oi + ei \quad (2)$$

$$Yi = b0 + b1Bi + b2Oi + b3BiOi + ei \quad (3)$$

where:

Yi = Job satisfaction.
Bi = Budget emphasis.
Oi = Organizational commitment.
ei = Error term.

High Budgetary Participation Situations

Table 3 presents the results of the main effects model Eq. (2) and the two-way interaction between budget emphasis and organizational commitment model Eq. (3) for the *high* budgetary participation subsample. Coefficient b_2 for the main effect Eq. (2) of organizational commitment on job satisfaction is, as expected, positive and highly significant (est. = 0.343; $p < 0.005$). The R^2 is 10.3%. Hypothesis H2, which states that organizational commitment is positively associated with job satisfaction in high budgetary participation situations, is supported. Note that coefficient b_3 for the two-way interaction Eq. (3) between budget emphasis and organizational commitment is, as expected, not significant (est. = –0.006; $p < 0.4740$).

Table 3. Results of Regression of Job Satisfaction on Budget Emphasis and Organizational Commitment for High Budgetary Participation Sub-sample ($n = 66$).

		Equation (2)			Equation (3)		
Variable	Coeff	Est	*t*-value	*p*	Est	*t*-value	*p*
Constant	b_0	57.140	6.174	0.001	53.455	0.945	0.174
Budget emphasis (BE)	b_1	0.097	0.169	0.433	0.411	0.086	0.466
Organizational commitment (OC)	b_2	0.343	2.654	0.005	0.410	0.401	0.345
BE × OC	b_3				−0.006	−0.066	0.474
R^2			0.103			0.103	
Adjusted R^2			0.074			0.059	
F value			3.607			2.368	
$p \leq$			0.016			0.039	
Adjusted R^2 explained by interaction term						−1.5%	

Low Budgetary Participation Situations

Table 4 presents the results of the main effects model Eq. (2) and the two-way interaction between budget emphasis and organizational commitment model Eq. (3) for the *low* budgetary participation sub-sample. Coefficient b_3 for the two-way interaction Eq. (3) between budget emphasis and organizational commitment is, as expected, highly significant (est. = –0.218; $p < 0.0015$). The R^2 is 46.3% and the adjusted R^2 increases by 12.3%, from 30.2% Eq. (2) to 42.5% Eq. (3) with the introduction of the two-way interaction term. These results provide initial support for hypotheses 3 and 4.

Monotonicity Test for the Low Budgetary Participation Subsample

Since the two-way interaction term for the *low* budgetary participation subsample is significant, additional tests are undertaken to ascertain the nature of this significant interaction. Schoonhoven (1981) argued that a monotonicity test is needed to confirm whether the relationship between the variables is non-monotonic in nature. He suggested that the graphing of the partial derivative equation is important in testing for non-monotonic effects. In order to test for such effects, the partial derivative of Eq. (3) over the range of budget emphasis needs to be computed. If the partial derivative graph intersects the organizational commitment axis (x-axis), the relationship between job satisfaction and budget emphasis is regarded as non-monotonic.

The partial derivative of Eq. (3) for the low budgetary participation subsample yields the following equation: Yi/Bi = b1+b3Oi. Substituting b1 = 10.653 and

Table 4. Results of Regression of Job Satisfaction on Budget Emphasis and Organizational Commitment for Low Budgetary Participation Sub-sample ($n = 46$).

		Equation (2)			Equation (3)		
Variable	Coeff	Est	*t*-value	*p*	Est	*t*-value	*p*
Constant	b_0	34.263	3.385	0.0010	−87.665	−2.232	0.0155
Budget emphasis (BE)	b_1	0.544	0.793	0.2160	10.653	3.302	0.0010
Organizational commitment (OC)	b_2	0.596	4.570	0.0001	3.222	3.879	0.0000
BE × OC	b_3				−0.218	−3.194	0.0015
R^2			0.333			0.463	
Adjusted R^2			0.302			0.425	
F value			10.727			12.081	
$p \leq$			0.000			0.000	
Adjusted R^2 explained by interaction term						12.3%	

b3 = –0.218 under Eq. (3) of Table 4, Yi/Bi = 10.653 + 0.218Oi. The point of inflection from the equation is 48.87 (10.653/0.218) when equating the partial derivative equation above to zero. As this point lies within the observed range of organizational commitment (13 to 63, Table 1) and is close to the mean of 48.84, budget emphasis has a non-monotonic effect on job satisfaction. This means that the effect of budget emphasis on job satisfaction is positive for organizational commitment scores below 48.87, but negative for organizational scores above 48.87. These effects are graphically shown in Fig. 2.

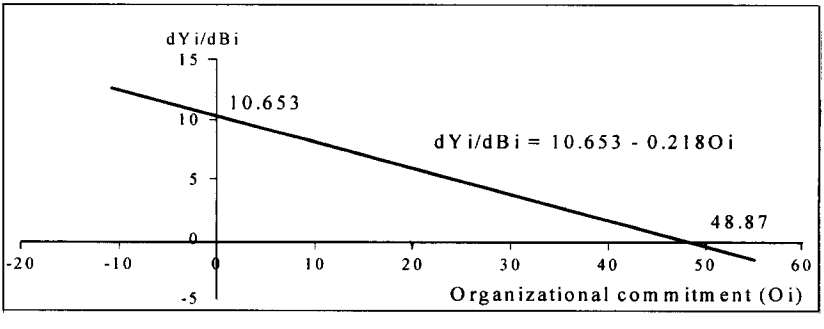


Fig. 2. Graphical Representation of the Relationship Between Budget Emphasis, Organizational Commitment and Job Satisfaction (Low Budgetary Participation Sub-sample).

Further Analysis of the Three-Way Interaction Model

To assist in the interpretation of the results presented in Tables 3 and 4 for Eq. (3), graphs of regression lines were plotted and presented in Fig. 3 for the high budgetary participation subsample, and Fig. 4 for the low budgetary participation subsample. Each graph presents two regression lines, one for high budget emphasis, based on the regression equation for high budget emphasis, and another for low budget emphasis, based on the regression equation for low budget emphasis. These regression lines indicate graphically how the relationship between organizational commitment and job satisfaction differs between high and low budget emphasis.

High budgetary participation situations. Figure 3 presents the results for the *high* budgetary participation subsample. It indicates that the slope of the high

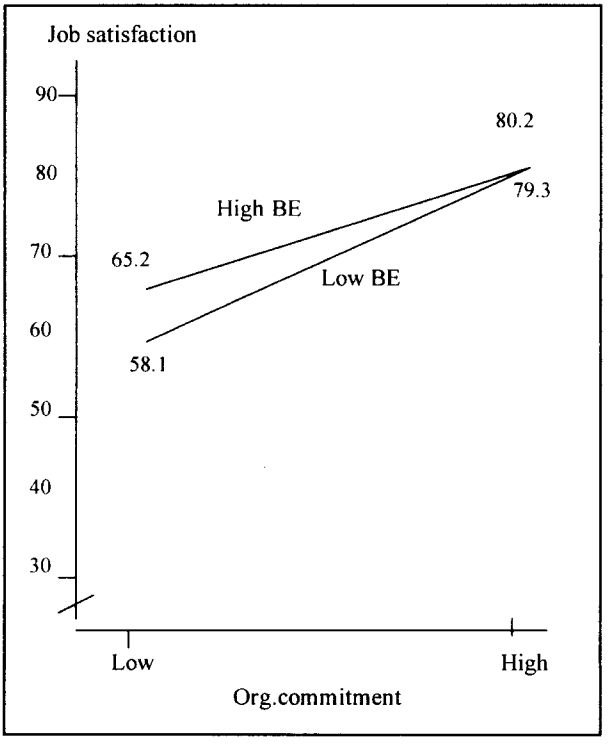


Fig. 3. Two-Way Interaction Between Budget Emphasis and Organizational Commitment Affecting Performance (High Budgetary Participation Subsample; $n = 66$).

budget emphasis regression line is not substantially different from that for the low budget emphasis regression line. This is supported by additional computation (Cohen & Cohen, 1983, p. 111) which indicates that the two slopes are not statistically significantly different ($p < 0.33$). This is consistent with the results in Table 3, which indicate that coefficient b_3 for the two way interaction between budget emphasis and organizational commitment is not significant ($p < 0.474$). This means that the relationship between organizational commitment and job satisfaction does not differ significantly between high budget emphasis and low budget emphasis. It also means that job satisfaction is high when organizational commitment is high and low when organizational commitment is low for both low and high budget emphasis. These results are in accordance with hypothesis H2, which states that in high budgetary participation situations, organizational commitment is positively associated only with job satisfaction, regardless of the levels of budget emphasis.

Low budgetary participation situations. For the low budgetary participation subsample, Table 4 indicates that coefficient b_3 for the two-way interaction between budget emphasis and organizational commitment is highly significant ($p < 0.0015$). Since this coefficient is significant only if the slopes of the two regression lines are significantly different (Cohen & Cohen, 1983, p. 316), the significant result found in Table 4 for coefficient b_3 provides the statistical support that the slopes of the two regression lines in Fig. 4 are significantly different. This is supported by additional computation (Cohen & Cohen, 1983, p. 111) which indicates that the slopes of the two regression lines are statistically significantly different ($p < 0.015$).

Note that Table 4 indicates that organizational commitment has a highly significant main effect on job satisfaction ($p < 0.001$ in Eq. (2)). However, it also indicates that *over and above* this main effect, there is a significant interactive effect. This is demonstrated by the substantial difference in the steepness of the slope between the two regression lines in Fig. 4. The slope for low budget emphasis is much steeper than that for high budget emphasis. In other words, the relationship between organizational commitment and job satisfaction is less positive for high budget emphasis than for low budget emphasis. This indicates the existence of a *negative* interaction and is consistent with the results in Table 4, which indicates that the two-way interaction coefficient b_3 is both significant and *negative* (est. = –0.218; $p < 0.0015$).

These results indicate that when organizational commitment is low, high budget emphasis is associated with higher job satisfaction compared with low budget emphasis in a low participatory environment. These results are not consistent with those found by Brownell (1982) and Brownell and Dunk (1991),

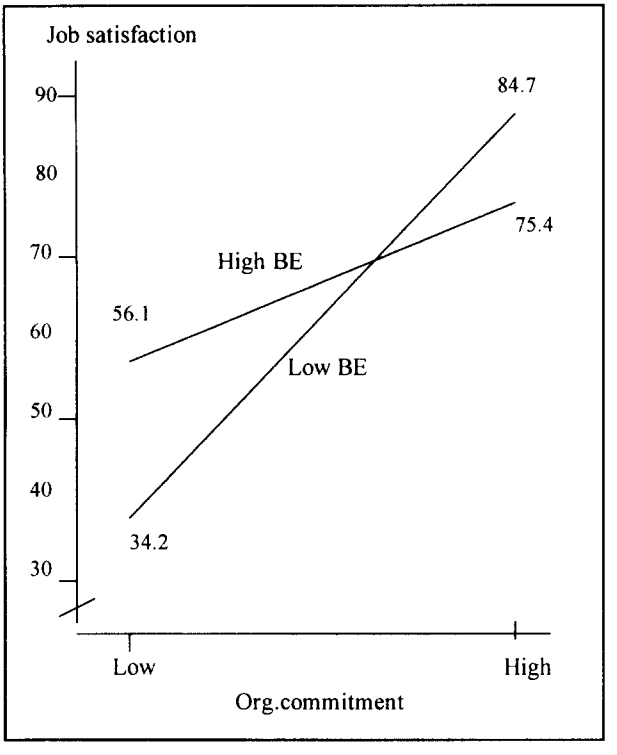


Fig. 4. Two-Way Interaction Between Budget Emphasis and Organizational Commitment Affecting Performance (Low Budgetary Participation Subsample; $n = 46$).

which indicate that in a low participatory environment, a low budget emphasis is associated with better behavioral outcomes than a high budget emphasis. However, these results are in accordance with those found by Lau et al. (1995). They are also consistent with hypothesis H3, which states that *when organizational commitment is low*, high budget emphasis is associated with higher job satisfaction than low budget emphasis. This indicates the importance of organizational commitment in the relationships among budget emphasis, participation and job satisfaction.

Figure 4 also indicates that when organizational commitment is high, low budget emphasis is associated with higher job satisfaction compared with high budget emphasis. These results are in accordance with hypothesis H4, which

states that *when organizational commitment is high*, low budget emphasis is associated with higher job satisfaction than high budget emphasis. They are also consistent with those found by prior studies (eg. Brownell, 1982; Brownell & Dunk, 1991), which indicate that in a low participatory environment, a low budget emphasis is associated with better behavioral outcomes than a high budget emphasis. Overall, the results in Figure 4 therefore indicate that prior studies' results (Brownell, 1982, Brownell & Dunk, 1991) on the beneficial effects of low budget emphasis in low participatory environment, are supported for the highly committed managers, but not for the lowly committed managers.

CONCLUSION

This study examines the interaction between budget emphasis, budgetary participation and organizational commitment affecting job satisfaction. The main motivation of this study was to extend the area of research on superior evaluative styles by investigating the moderating effects of organizational commitment as a possible explanation for the inconsistencies in the research results of this research area. Specifically, the results of Brownell and Dunk (1991) were not consistent with those of Lau et al. (1995) with respect to the effects of the *low* budget emphasis and *low* participation combination on managers' behavior.

In general, the results support the hypotheses developed in this study. The major findings are as follows. A statistically significant three-way interaction was found between budget emphasis, budgetary participation and organizational commitment affecting job satisfaction. There was also a significant interaction between budget emphasis and organizational commitment affecting job satisfaction for the low budgetary participation situations. These results support the expectation that in *low* budgetary participation situations: (i) *high* budget emphasis is associated with higher job satisfaction than low budget emphasis for *lowly* committed managers; and (ii) *low* budget emphasis is associated with higher job satisfaction than high budget emphasis for *highly* committed managers. Finally, a significant positive relationship between organizational commitment and job satisfaction was found for both the high budgetary participation situations and the low budgetary participation situations.

These results provide interesting and important insights into the complex relationships among budget emphasis, budgetary participation, organizational commitment and managers' job satisfaction. They may have important theoretical and practical implications. With respect to theory, the results of this study may provide some degree of explanation and therefore may assist in the resolution of the conflicting results of Brownell and Dunk (1995) and Lau et

al. (1995). The omission of the moderating effects of organizational commitment on the interaction between budget emphasis and budgetary participation may partly be the reason why the results of these two studies differ. The resolution of such discrepancy in research findings may be an important contribution to the further development of theory in this important research area.

From a practical perspective, the results of this study from the financial services sector may provide some insights into the reactions of managers to control systems in general, and to the control systems in the financial services sector in particular. Whilst there has been considerable research in the areas of performance evaluative style and budgetary participation, it appears that the evidence has been accumulated predominantly from the manufacturing sector (eg. Brownell, 1982; Brownell & Hirst, 1986; Brownell & Dunk, 1991; Nouri, 1994; Lau et al., 1995; Nouri & Parker, 1996), and to a lesser extent, from the merchandising sector (Harrison, 1992, 1993). There is a dearth of research evidence from the services sector in general and the financial services sector in particular (Pope & Otley, 1996; Lau & Tan, 1998). Hence, it remains unclear if the evidence accumulated from the manufacturing and merchandising sectors could be generalized to the services sector. The results of this study from the financial services sector may therefore be important to system designers who are involved in the design and implementation of control systems in the financial services sector. As this sector is usually a major sector in most economies, and as it may have features which are unique (Mester, 1992; Calem & Rizzo, 1992; Lau & Tan, 1998; Saunders & Lange, 2000), results from studies in this sector may provide evidence which system designers could rely upon with greater confidence than with evidence collected from other sectors.

As with most studies, there are limitations associated with this study. First, this study was undertaken with subjects drawn solely from the financial services sector. Hence, generalizing the results to other industry sectors should be undertaken with caution. Second, even generalizing the results of this study to the financial services sector should be undertaken with caution since, as mentioned previously, there are very few studies in this research area that were undertaken in the financial services sector. Further studies are therefore needed to verify the findings in order to increase their generalizability. This study has also not control for organizational culture that may vary across organizations. In addition, following findings by Lincoln and Kallenberg (1990) that national culture could have an impact on the managers' level of organizational commitment and job satisfaction, this potential avenue for research should not be overlooked. Finally, even though the questionnaire survey method has been widely used in this research area (e.g. Hopwood, 1972; Brownell, 1982, 1985; Shields & Shields, 1998; Oley & Pollanen, 2000), and according to Brownell

(1995, p. 55), "easily the most popular research method in empirical work in management accounting", like other research methods, it has its strengths and limitations (Marsh, 1982). Hence, whilst its strength includes the enhancement of external validity, its limitations may include the potential for measurement error in the instruments, leniency error, acquiescence error, halo error, range restriction and internal validity threats. Since methodological pluralism in research is desirable, future studies could employ other methods such as the experimental method, field studies and in-depth case studies to further investigate the complex relationships in this research area.

Nevertheless, despite these limitations, this study developed and empirically tested a plausible model to explain the relationships among evaluative style, budgetary participation, organizational commitment and job satisfaction, an important and complex area of management accounting research. The general support for the hypotheses developed in this study provides credibility to the theory developed and opportunities for additional research. Such developments may assist in the resolution of discrepancies in research findings, in the further development of theories and in the design of effective control systems in the important financial services sector and other sectors.

REFERENCES

Angle, H. L., & Perry, J. L. (1981). An Empirical Assessment of Organizational Commitment and Organizational Effectiveness. *Administrative Science Quarterly*, *26*(1), 1–14.

Bateman, T., & Strasser, S. (1984). A Longitudinal Analysis of the Antecedents of Organizational Commitment. *Academy of Management Journal*, *27*, 95–112.

Blau, G. J. (1987). Using a Person-Environment Fit Model to Predict Job Involvement and Organizational Commitment. *Journal of Vocational Behavior*, *30*(3), 240–257.

Briers, M. L., & Hirst, M. K. (1990). The Role of Budgetary Information in Performance Evaluation. *Accounting, Organizations and Society*, 373–398.

Brownell, P. (1982). The Role of Accounting Data in Performance Evaluation, Budgetary Participation, and Organizational Effectiveness. *Journal of Accounting Research*, *20*(1), 12–27.

Brownell, P. (1985). Budgetary Systems and the Control of Functionally Differentiated Organizational Activities. *Journal of Accounting Research*, *23*(2), 502–512.

Brownell, P. (1995). *Research methods in management accounting*, Victorian Printing, Blackburn.

Brownell, P., & Dunk, A. S. (1991). Task Uncertainty and Its Interaction with Budgetary Participation and Budget Emphasis: Some Methodological Issues and Empirical Investigation. *Accounting, Organizations and Society*, *16*(8), 693–703.

Brownell, P., & Hirst, M. (1986). Reliance on Accounting Information, Budgetary Participation, and Task Uncertainty: Tests of a Three-way Interaction. *Journal of Accounting Research*, *24*(2), 241–249.

Butler, J. K. (1983). Value Importance as a Moderator of the Value Fulfillment Job Satisfaction Relationship: Group Differences. *Journal of Applied Psychology*, *68*(3), 420–428.

Calem P. S., & Rizzo, J. A. (1992). Banks As Information Specialists: The Case of Hospital Lending, *Journal of Banking and Finance*, 1123–1141.

Campbell, T., & Kracaw, W. (1980). Information Production, Market Signalling and the Theory of Financial Intermediation, *Journal of Finance*, *35*, 863–882.

Cherrington, D. J., & Cherrington, J. O. (1973). Appropriate Reinforcement Contingencies in the Budgeting Process: Empirical Research in Accounting: Selected Studies. *Journal of Accounting Research*, (Supplement)(11), 225–253.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral Sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.

Collins, F. (1978). The Interaction of Budget Characteristics and Personality Variables with Budgetary Response Attitudes. *The Accounting Review*, *53*(2), 324–335.

DeCotiis, T. A., & Summers, T. P. (1987). A Path Analysis of a Model of the Antecedents and Consequences of Organizational Commitment, *Human Relations*, *40*, 445–470.

Dunham, R. B., Smith, F. J., & Blackburn, R. S. (1977). Validation of the Index of Organizational Reactions with the JDI, the MSQ, and Faces Scales. *Academy of Management Journal*, *20*(3), 420–432.

Frucot, V., & Shearon, W. T. (1991). Budgetary Participation, Locus of Control, and Mexican Managerial Performance and Job Satisfaction. *The Accounting Review*, *66*(1), 80–99.

Govindarajan, V. (1984). Appropriateness of Accounting Data in Performance Evaluation: An Empirical Examination of Environmental Uncertainty as an Intervening Variable. *Accounting, Organizations and Society*, *9*(2), 125–135.

Harrison, G. L. (1992). The Cross-Cultural Generalizability of the Relation Between Participation, Budget Emphasis and Job Related Attitudes. *Accounting, Organizations and Society*, *17*(1), 1–15.

Harrison, G. L. (1993). Reliance on Accounting Performance Measures in Superior Evaluative Style: The Influence of National Culture and Personality. *Accounting, Organizations and Society*, *18*(4), 319–339.

Hopwood, A. G. (1972). An Empirical Study of the Role of Accounting Data in Performance Evaluation. *Journal of Accounting Research*, *10*(Supplement), 156–182.

Horngren, C. T., Foster, G., & Datar, S. M. (2000). *Cost accounting: A managerial emphasis* (10th ed.). Prentice Hall. New Jersey.

Kaplan R. S. (1983). Measuring Manufacturing Performance: A New Challenge for Managerial Accounting Research. *The Accounting Review*, 686–705.

Kaplan R. S., & Norton, D. P. (1996). Using the Balanced Scorecard as a Strategic Management System. *Harvard Business Review*, 75–85.

Kren, R. L., & Liao, W. M. (1988). The Role of Accounting Information in the Control of Organizations: A Review of the Evidence. *Journal of Accounting Literature*, 280–309.

Kompass Australia 1998 (1998). Prahran, Victoria: Peter Isaacson Publications.

Lau, C. M., Low, L. C., & Eggleton, I. R. C. (1995). The Impact of Reliance on Accounting Performance Measures on Job-Related Tension and Managerial Performance: Additional Evidence. *Accounting, Organizations and Society*, *20*(5), 359–381.

Lau, C. M., & Tan, J. J. (1998). The Impact of Budget Emphasis, Participation and Task Difficulty: A Cross-Cultural Study of the Financial Services Sector. *Management Accounting Research*, *9*, 163–193.

Lindsay, R. M., & Ehrenberg, S. C. (1993). The Design of Replicated Studies, *The American Statistician*, 217–228.

Lincoln, J. R., & Kallenberg, A. L. (1990). *Culture, control, and commitment*. Cambridge: Cambridge University Press.

Locke, E. A. (1976). The Nature and Causes of Job Satisfaction. In: M. D. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology* (pp. 1297–1349). Chicago: Rand McNally College Publishing Company.

Lowe, E. A., & Shaw, R. W. (1968). An Analysis of Managerial Biasing: Evidence from a Company's Budgeting Process. *Journal of Management Studies*, *5*(October), 304–315.

Marsh, C. (1982). *The survey method: The contribution of surveys to sociological explanation*, London: George Allen and Unwin.

Mathieu, J. E., & Zajac, D. M. (1990). A Review and Meta-Analysis of the Antecedents, Correlates, and Consequences of Organizational Commitment. *Psychological Bulletin*, *108*(2), 171–194.

Merchant, K. A. (1981). The Design of the Corporate Budgeting System: Influences on Managerial Behavior and Performance. *The Accounting Review*, *56*(4), 813–829.

Mester, L. J. (1992). Traditional and Non-Traditional Banking: An Information-Theoretic Approach, *Journal of Banking and Finance*, 545–566.

Meyer, J. P., Paunonen, S. V., Gellatly, I. R., Goffin, R. D., & Jackson, D. N. (1989). Organizational Commitment and Job Performance: It is the Nature of the Commitment That Counts. *Journal of Applied Psychology*, *74*(1), 152–156.

Meyer, J. P., Allen, N. J., & Gellatly, I. R. (1990). Affective and Continuance Commitment to the Organization: Evaluation of Measures and Analysis of Concurrent and Time-Lagged Relations. *Journal of Applied Psychology*, *75*(6), 710–720.

Meyer, J. P., & Allen, N. J. (1991). A Three-Component Conceptualization of Organizational Commitment. *Human Resource Management Review*, *1*(1), 61–89.

Milani, K. (1975). The Relationship of Participation in Budget-Setting to Industrial Supervisor Performance and Attitudes: A Field Study. *The Accounting Review*, *50*(2), 274–284.

Mowday, R. T., Porter, L. W., & Steers, R. M. (1982). *Employee-organization linkages: The psychology of commitment, absenteeism, and turnover*. New York: Academic Press.

Mowday, R. T., Steers, R. M., & Porter, L. W. (1979). The Measurement of Organizational Commitment. *Journal of Vocational Behavior*, *14*(2), 224–247.

Nouri, H. (1994). Using Organizational Commitment and Job Involvement to Predict Budgetary Slack: A Research Note. *Accounting, Organizations and Society*, *19*(3), 289–295.

Nouri, H., & Parker, R. J. (1996). The Effect of Organizational Commitment on the Relation Between Budgetary Participation and Budgetary Slack. *Behavioral Research in Accounting*, *8*, 74–90.

Nouri, H., & Parker, R. J. (1998). The Relationship Between Budget Participation and Job Performance: The Roles of Budget Adequacy and Organizational Commitment. *Accounting, Organizations and Society*, *23*(5/6), 467–483.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement*. London: Pinter Publishers.

Otley, D. T. (1978). Budget Use and Managerial Performance. *Journal of Accounting Research*, *16*(1), 122–149.

Otley, D. T., & Fakiolas, A. (2000). Reliance on Accounting Performance Measures: Dead end or New Beginning. *Accounting, Organizations and Society*, 497–510.

Otley, D. T., & Pollanen, R. M. (2000) Budgetary Criteria in Performance Evaluation: A Critical Appraisal Using New Evidence. *Accounting, Organizations and Society*, 483–496.

Porter, L., W., Steers, R. M., Mowday, R. T., & Boulian, P. V. (1974). Organizational Commitment, Job Satisfaction, and Turnover Among Psychiatric Technicians. *Journal of Applied Psychology*, *59*(5), 603–609.

Pope, P., & Otley, D. T. (1996). Budgetary control and performance evaluation: an empirical analysis of bank branches. Working Paper, Lancaster University.

Poznanski, P. J., & Bline, D. M. (1997). Using Structural Equation Modeling to Investigate the Causal Ordering of Job Satisfaction and Organizational Commitment Among Staff Accountants. *Behavioral Research in Accounting*, 154–171.

Randall, D. M. (1990). The Consequences of Organizational Commitment: Methodological Investigation. *Journal of Organizational Behavior, 11*(5), 361–378.

Ross, A. (1994). Trust as a Moderator of the Effect of Performance Evaluation Style on Job Related Tension: A Research Note, *Accounting, Organizations and Society*, 629–635.

Saunders, A., & Lange, H. (2000). *Financial institutions management; A modern perspective.* McGraw-hill Roseville.

Scarpello, V., & Campbell, J. P. (1983). Job Satisfaction: Are all the Parts There?. *Personnel Psychology, 36*(3), 577–600.

Schoonhoven, C. B. (1981). Problems with Contingency Theory: Testing Assumptions Hidden with the Language of Contingency Theory. *Administrative Science Quarterly, 26*(3), 349–377.

Shields, J. F., & Shields, M. D. (1998). Antecedents of Participative Budgeting, *Accounting, Organizations and Society*, 49–76.

Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement.* Chicago: Rand McNally.

Sommer, S. M., Bae, S. H., & Luthans, F. (1996). Organizational Commitment Across Cultures: The Impact of Antecedents on Korean Companies. *Human Relations, 49*(7), 977–993.

Steer, R. M. (1977), Antecedents and Outcomes of Organizational Commitment, *Administrative Science Quarterly, 22*, 46–56.

Weiss, D. J., Dawis, R. V., England, G. W., & Lofquist, L. H. (1967). *Manual for the Minnesota Satisfaction Questionnaire. Minnesota Studies in Vocational Rehabilitation*, 22. Minneapolis: University of Minnesota, Industrial Relations Center, Work Adjustment Project.

White Pages (1997/98). Telstra Corporation Limited.

THE EFFECTS OF RED-FLAG ITEMS, UNFAVORABLE PROJECTION ERRORS, AND TIME PRESSURE ON TAX PREPARERS' AGGRESSIVENESS

Richard I. Newmark and Khondkar E. Karim

ABSTRACT

This paper examines how a red-flag item (RFI), an unfavorable projection error (UPE) and time pressure affect the aggressiveness of tax preparers' recommendations. We hypothesize that a RFI causes preparers to make more conservative recommendations, and UPE and time pressure lead to more aggressive recommendations. The hypotheses were tested on 153 tax practitioners using a hypothetical tax scenario. There was a strong indication that practitioners made more conservative recommendations when a RFI was present. Additionally, the results indicated that the effect of time pressure on tax preparers' recommendations was dependent upon other variables in their decision-making environment.

INTRODUCTION

Tax professionals resolve issues for their clients on a daily basis. Deciding many of these issues is fairly straightforward because the law is clear and applies to facts unambiguously to support a particular position. Tax preparers experience

Advances in Accounting Behavioral Research, Volume 5, pages 213–243.

ISBN: 0-7623-0953-9

difficulty, however, when the law does not clearly support a particular position because preparers must make recommendations to clients who are primarily interested in black-and-white answers to their tax issues (Raabe et al., 1997). When resolving ambiguous issues, both issue- and non-issue-relevant information has been shown to influence tax preparers' decisions. For example, preparers made more conservative recommendations when audit probability was high (Kaplan et al., 1988; McGill, 1988) and more aggressive recommendations when they perceived a loss situation (Schisler, 1994).

Although tax preparers may experience acute time pressure during tax season, the effects of audit probability and loss frame of reference (or any other previously tested variables) on resolving ambiguous tax issues have not been previously studied under time pressure. Prior time pressure studies in auditing (Brown & Solomon, 1992; Choo & Firth, 1993) and tax (Spilker, 1995; Spilker & Prawitt, 1997) indicate that participants simplify their decision task under such pressure. Tax preparers may simplify their decision under time pressure by resolving ambiguous issues in the client's favor because they are supposed to be taxpayer advocates (AICPA, 1997, TX §112). Additionally, resolving ambiguous issues in a client's favor rarely subjects either the preparer or the client to IRS penalties (IRC, §§6694 and 6662).[1]

The purpose of this study is to examine how non-issue-relevant information – a red-flag item (i.e. a tax return item perceived to increase audit probability) and an unfavorable projection error (i.e. an error in a year-end tax projection that causes the preparer to underestimate the client's tax liability) – affect preparers' recommendations to clients on an ambiguous tax issue, and whether or not preparers' recommendations will differ based on time pressure. If non-issue-relevant information affects tax preparers' confidence in a pro-client position, it could increase their clients' total tax-related costs, as overconfidence could result in additional cost from an unexpected audit, while underconfidence may cause the preparer to recommend an unnecessarily conservative position (resulting in a larger-than-necessary tax liability). Additionally, understanding how time pressure affects preparers' recommendations is important because time pressure can be controlled (at least to some extent) by managing preparers' workloads.

This study extends tax preparer research on IRS audit probability by examining the effect of a "real-world" item believed to attract IRS scrutiny on tax preparers' recommendations. Prior studies explicitly stated audit probability (e.g. Kaplan et al., 1988, McGill, 1988). This research also suggests that time pressure can affect the judgment of tax professionals on an ambiguous tax issue, an issue that has not been examined in prior tax research. These contributions are discussed in greater detail at the conclusion of the paper.

The remainder of this paper is divided into four sections. The first section describes the research hypotheses, including prior research that serves as a foundation for the hypotheses. The second section offers the research methodology, and the third section presents the results. The final section summarizes and discusses the research findings, contributions, and limitations.

PRIOR RESEARCH AND HYPOTHESES

Heuristic cues are variables not directly related to the tax issue that can be processed with little cognitive effort by applying simple decision rules (*decision heuristics*) (e.g. Chaiken, 1980; Chaiken, et al., 1989). Several heuristic cues have been shown to influence the resolution of ambiguous issues by tax professionals, including client risk-preference (Duncan et al., 1989; Schisler, 1994), client importance (McGill, 1988; Reckers et al., 1991), and explicitly stated audit probability (McGill, 1988; Kaplan et al., 1988).[2] However, these cues may not be available to tax preparers because they often prepare returns for clients they never meet, which limits access to client risk-preference and client importance,[3] and the IRS does not disclose its audit selection function (Raabe et al., 1997). Prior studies have not examined the effects of heuristic cues on preparers' decisions under time pressure. For these reasons, the heuristic cues included in this study – a red-flag item and unfavorable projection error – are accessible to tax preparers because they are typically included in clients' tax files (Kastantin, 1988). Moreover, since it is not known if time pressure affects the influence of conservative cues different from its effect on aggressive cues, this study hypothesizes a conservative heuristic cue (red-flag item) and an aggressive heuristic cue (unfavorable projection error).

Red-Flag Item

Deciding how to report an ambiguous item on a tax return is difficult since there is uncertainty about how the IRS will treat the issue upon examination. There is also uncertainty as to whether or not the IRS will even examine the ambiguous item. McGill (1988) defined the likelihood of IRS scrutiny, probability of detection, as the product of the probability that the tax return will be audited and the probability that the ambiguous item will be examined, given an audit. Therefore, the presence of a red-flag item increases the risk that the IRS will deny an aggressive position by increasing the likelihood that the tax return will be audited, which, in turn, increases the probability that the ambiguous item will be examined.

Although prior research has shown that tax preparers make more conservative judgments when audit probability is known and sufficiently high (Kaplan et al., 1988; McGill, 1988), tax preparers do not know audit probability with certainty. In practice, they make rough audit probability assessments based on anecdotal evidence, including the presence of red-flag items (items on the tax return that are believed to be carefully scrutinized by the IRS) (Wartzman, 1993). Kaplan et al. (1988) and Schisler (1994) made reference to the red-flag item and used it as the tax issue in question instead of being unrelated to the tax issue as in the present study. Moreover, in both studies, audit probability was explicitly stated, so the red-flag item was not used as an indirect indicator of audit probability.

Popular business periodicals and news shows have reinforced the idea that red-flag items increase the likelihood of IRS scrutiny, as these sources often report that the IRS routinely identifies red-flag items, including travel and entertainment expenses, business auto expenses, casualty losses, barter income, hobby losses, tax shelter investments, and home office deductions (Flanagan, 1981; Wiltsee, 1984; Wartzman, 1993; Jenkins & Schuch, 1997). Since tax preparers believe that red-flag items are closely monitored by the IRS (Flanagan, 1981; Wiltsee, 1984; Wartzman, 1993; Jenkins & Schuch, 1997), the presence of such items should increase perceived audit probability (the probability that the IRS will scrutinize the tax return). Therefore, tax preparers are expected to take conservative positions on ambiguous issues when red-flag items are present because an audit increases the likelihood that the IRS will examine and deny an aggressive position. This leads to the first research hypothesis, which is stated in alternative form.

H1: Tax preparers will make more conservative recommendations to clients on ambiguous tax issues when at least one red-flag item is present than when no red-flag item is present.

Unfavorable Projection Error (UPE)

Tax professionals often project a client's tax liability at year-end meetings (usually during November or December) to determine what can be done to further reduce the client's tax liability (Kastantin, 1988) and prevent penalties (e.g. underpayment of estimated taxes), especially when the tax law significantly changes during the year (Raby, 1974). After projecting the client's tax liability and ascertaining his/her current-year tax payments (including additional payments based on the year-end projection), the preparer can inform the client of his/her projected refund or tax due months before the tax return is actually filed.

Later, when preparing the actual tax return, preparers (or supervisors) may find errors in the year-end projection that cause the actual tax liability to exceed the projected liability (an unfavorable projection error, or UPE). Informing a client of a substantially larger tax liability than expected can cause him/her to become upset, especially if he/she does not have funds available to pay the additional tax (Rachlin, 1983). A dissatisfied client can cause the preparer (or the supervisor) immediate negative personal and/or financial consequences. For example, an irate client may chastise the preparer (or supervisor), or cease doing business with the preparer's accounting firm. Further, the tax preparer may receive some form of punishment from within the accounting firm for harming the firm's reputation and profitability.

A tax preparer (or supervisor) can lessen the immediate negative consequences of a UPE by recommending that the client take an aggressive position on an ambiguous tax issue, provided that the projection was based on taking a conservative (or less aggressive) position, and the tax reduction from taking a more aggressive position must partially or completely offset the UPE. By offsetting the UPE, the client will not be unpleasantly "surprised" by an unexpectedly large tax liability, thereby reducing negative repercussions from the client. Of course, the client should be informed and agree to the aggressive position.

The primary risk to the tax preparer is that the client's tax return may be audited. However, the preparer (or supervisor) does not have to worry about that possibility for at least a year, which is the amount of time it usually takes for the IRS to notify a taxpayer that his/her tax return will be audited (Lohse, 1994). Even if the tax return is audited, the IRS may allow the aggressive tax position. Thus, taking an aggressive position on an ambiguous issue replaces the (almost) certain immediate negative consequences from making a UPE with a possibility of severe negative consequences in the future.

Taking an action to avoid certain and immediate negative consequences in exchange for a chance to receive no (or minimal) immediate negative consequences, and only a low possibility of future negative consequences, is consistent with Thorndike's (1913) well known Law of Effect.[4] That is, people tend to avoid actions that result in discomfort or punishment and repeat actions that result in pleasure or reward.

However, one may refrain from engaging in an activity to avoid punishment (e.g. recommend an aggressive position to a client in order to avoid a UPE if the preparer previously recommended the conservative position to that client) if that activity is believed to be unethical. Jones (1991) believed that the likelihood of making an ethical decision was positively related to the issue's *moral intensity* – to the extent of its issue-related moral imperative (Jones, 1991, p. 372). This concept has been empirically supported in the marketing ethics literature

(e.g. Singhapaki et al., 1999). That is, tax preparers would be more likely to make an aggressive recommendation and avoid punishment (consistent with the Law of Effect) if they perceive the choice to be low in moral intensity and make the conservative recommendation (contrary to the Law of Effect) if they believe the choice is high in moral intensity.

Moral intensity has six dimensions: *magnitude of consequences* (sum of benefits and harms to the client), *probability of effect* (probability that the act will take place and probability of it causing harm [benefit]), *temporal immediacy* (time from the present to the onset of consequences), *social consensus* (beliefs of peers, co-workers, etc. that the decision is unethical), *proximity* (nearness [social, cultural, psychological, physical] of preparer to the client), and *concentration of effect* (inverse function of number of people affected by an act of a given magnitude). The *magnitude of the consequences* of recommending an aggressive position in light of the UPE do not seem to be overly negative. The biggest "harm" that the client would encounter is an IRS audit, which might lead to the same tax liability as if the conservative position were initially taken plus interest on the unpaid liability. However, the *probability of effect* (probability of an audit, and probability that the aggressive position would be denied) is uncertain. First, the tax return must be "flagged" for an audit, which depends on the nature and dollar amount of the item (e.g. is it an area subject to abuse?). Even if the item is audited, the most serious consequence (increasing the liability to reflect the conservative position plus interest on the unpaid balance) will only occur if the IRS is successful in denying the aggressive position on an issue that can go either way. Moreover, the *temporal immediacy* is low as it usually takes at least a year for the IRS to notify the client that they wish to examine his/her tax return. Additionally, the *social consensus* is likely to be low as preventing negative consequences from a UPE is in the best interest of the accounting firm, as well as the individual making the decision (and would not be considered illegal under IRC §§6694 and 6662 or a violation of the AICPA standards of behavior (AICPA, 1997, TX §112). Therefore, the moral intensity of recommending an aggressive position as a result of a UPE appears to be low, and thus, would likely not discourage a preparer from recommending an aggressive position to minimize the immediate negative consequences of a UPE.

These predicted tax preparer reactions to UPE are also consistent with Prospect Theory (Kahneman & Tversky, 1979; Schisler, 1994), which states that people are risk-seeking when they view a situation as a loss (i.e. a UPE). Meaning, individuals prefer taking a risky position that is uncertain (i.e. taking an aggressive tax position) rather than taking action that results in a certain loss (i.e. taking a conservative tax position). Wright and Weitz (1977) provided additional support for making an aggressive recommendation in this situation, as they found that

decision makers selected more risky choices when the outcome horizon is distant than when it is imminent. Jones (1991) proposed an issue-contingent model containing a new set of variables called "moral intensity" (using concepts, theory, and evidence derived largely from social psychology) and he argues that moral intensity influences every component of moral decision making and behavior. Therefore, tax preparers may take an aggressive position on ambiguous issues that are not high in moral intensity to avoid the negative consequences associated with making an unfavorable projection error. This leads to the second hypothesis, which is stated in alternative form:

> *H2*: Tax preparers will make more aggressive recommendations to clients on ambiguous tax issues when they are responsible for an unfavorable projection error than when no unfavorable projection error was made.

Time Pressure

Early research concerning the effects of time pressure on tax preparers consisted of surveys indicating that time pressure causes a majority of tax accountants (and other public accountants) to underreport chargeable hours (Douglas, 1979; Lightner et al., 1982; Smith & Hutton; 1995).[5] Two more recent studies focused on the effects of time pressure on tax practitioners' ability to select relevant key words when researching a particular issue (Spilker, 1995; Spilker & Prawitt, 1997). In Spilker (1995), the experienced tax professionals who had both declarative knowledge (specific knowledge of facts and concepts in a given area) and procedural knowledge (how to perform tax research) chose significantly more relevant key words when under moderate time pressure than when time pressure was absent, but graduate tax students with only declarative knowledge showed no significant change. This is consistent with Spilker and Prawitt (1997), who showed that experienced tax researchers under moderate time pressure increased the proportion of time spent on searching salient areas of the tax research database to a greater extent than did graduate tax students.

These studies demonstrated that time pressure can affect tax preparers' decisions in a complex task, but they lacked many elements typically present when having to make a recommendation to a client. Moreover, no single time-pressure study contained all the elements of interest in this study: a complex task, accounting participants, ambiguous information, and non-issue-relevant information.[6] Therefore, time pressure studies from other disciplines which contain some of the elements relevant to the current study will be used to develop the time-pressure hypothesis.

When solving complex tasks under time pressure, high time-pressure participants used simpler, less time consuming decision rules than did participants

under low time-pressure (Wright, 1974; Christensen-Szalanski, 1980; Rothstein, 1986). For example, in Wright's study, low time-pressure participants used a compensatory decision-making strategy when judging 30 automobiles on five criteria, but participants under high time-pressure focused only on negative information. Auditors also used simpler decision strategies to solve complex tasks under time pressure (Brown & Solomon, 1992; Choo & Firth, 1993). In both Brown and Solomon (1992) and Choo and Firth (1993), audit seniors were asked to complete an audit task that required them to use a compensatory decision strategy in which an abundance of one attribute is allowed to compensate for a lack of another attribute. Auditors in the low time-pressure treatment used the appropriate compensatory decision strategy (the strategy that they were trained to use), but high time-pressure participants used a simpler strategy in spite of their training. Additionally, Kelley and Margheim (1990) found that practicing auditors under time budget pressure simplified their decision-making process by engaging in audit quality reduction acts (e.g. premature signoffs, accepting weak client explanations).[7] Since high time-pressure participants in these studies used simpler decision rules, it appears that auditors, like other decision makers, simplified their decisions when evaluating complex information under time pressure.

Although auditors and other decision makers used simplifying decision rules when evaluating complex information under time pressure, the nature of the judgment task may influence the effect of time pressure. An auditor's task is to evaluate the quality of a client's financial information. The likelihood that an auditor discovers an inadequacy in the information likely increases with additional time, which may not please the client, but reduces the risk of audit failure; whereas decreasing the amount of time available for an audit will tend to have the opposite effect. However, a tax professional is a client advocate. As such, increasing the amount of time to research a particular issue may increase the amount of pro-client support for the issue, which is beneficial to the client; whereas decreasing the amount of time researching an issue reduces the opportunity to find pro-client support for the issue. Thus, even though auditors and tax professionals in CPA firms are subject to strict time budgets and have similar educational backgrounds, time pressure may cause them to simplify their decisions in different ways.

Svenson and Edland's (1987) time-pressure study contained another element relevant to the current study – ambiguous information. Participants had to choose between two college apartments based on three criteria: traveling time to campus, size, and standard (old or modern). Other participants previously rated each apartment in the pair to be equally desirable. This decision is similar to the ambiguity tax professionals face when trying to resolve an issue that has

an equal amount of support for both the aggressive and conservative positions. Low time-pressure participants focused on apartment size while high time-pressure participants focused on traveling time to campus. In a related experiment that involved rating individual apartments, Svenson and Edland (1987) showed that high time-pressure participants rated individual apartments more negatively than low time pressure participants. These results provide additional evidence that people use different decision rules when considering ambiguous information under time pressure.

The heuristic-systematic model of persuasion (HSM) suggests that decision makers (e.g. tax professionals) who normally engage in systematic processing (thoughtful integration of all issue-relevant information in a cohesive manner to make a judgment [Chaiken et al., 1989]) will search for heuristic cues (non-issue-relevant information that can be easily processed by applying a simple decision rule) to help them resolve issues when an aspect of the decision-making environment (e.g. ambiguous information, time pressure) prevents them from reaching a conclusion based on systematic processing. Tax practitioner studies using ambiguous tax issues *without* time pressure showed a variety of heuristic cues influenced practitioners' judgments, such as client risk-preference (Duncan et al., 1989; Schisler, 1994), client importance (McGill, 1988; Reckers et al., 1991), and explicitly stated audit probability (Kaplan et al., 1988). Although Spilker (1995) and Spilker and Prawitt (1997) found that under moderate time pressure experienced tax professionals focused on the most issue-relevant information (a subset of all issue-relevant information), there were no available heuristic cues on which to focus.

When heuristic cues were available, Ratneshwar and Chaiken (1991) found that participants under low time pressure focused on issue-relevant information (information in a product advertisement) to determine the efficacy of a product while high time-pressure participants focused on a heuristic cue (i.e. source expertise). Anderson et al. (1994) used the HSM to show that auditors used source expertise to determine the reliability of a manager's assertions. Based on these studies, it appears that tax preparers may be influenced by heuristic cues when confronted with an ambiguous issue or when under time pressure.

In the current study, tax preparers were asked to make a recommendation to a client on an ambiguous tax issue while under time pressure. While ambiguous information causes tax preparers to search for heuristic cues to help them resolve an issue, time pressure limits the extent of this search. Therefore, tax preparers will likely be influenced by the most easily accessible cues.

Tax practitioners, unlike auditors, are client advocates. Clients pay preparers to resolve issues in the client's favor whenever the facts support taking such a position. In fact, failing to deduct an ambiguous item for which there is

reasonable support is contrary to a tax preparer's role as a taxpayer advocate (AICPA, 1997, TX §112) since resolving ambiguous issues in a client's favor is allowable under Rule 102 of the AICPA Code of Professional Conduct (Raabe et al., 1997, p. 12). Also, claiming the aggressive position does not violate §10.34(a)(1) (Standards for Advising with Respect to Tax Return Positions and for Preparing or Signing Returns) of the Treasury Department's Circular 230 (IRS, 1994) or subject the tax preparer to an accuracy-related penalty under IRC §6694, as the standard in both rules is that the position have at least a one-in-three possibility of prevailing on its merits. Moreover, a 50% likelihood that an ambiguous issue would be sustained on its merits is likely sufficient for the client to avoid penalties (IRC §6662). Thus, time pressure may cause tax preparers to focus on their role as client advocate, thereby making more aggressive recommendations on ambiguous tax issues. This leads to the third hypothesis, which is stated in alternative form:

H3: Tax preparers will make more aggressive recommendations to clients on ambiguous tax issues when under high time pressure than under low time pressure.

RESEARCH METHODOLOGY

Sample

Two-hundred forty self-administering test instruments were randomly distributed on computer diskettes to partners at local, regional, national, and Big-5 accounting firms. The partners then distributed the identically-labeled test instruments to qualified employees, defined in this study as accountants who spend at least 40% of their time performing tax-related work. One hundred fifty-three useable instruments were returned, for a response rate of 64%.[8] Table 1 summarizes participant demographic information for the sample. Cell means on continuous demographic variables (age, accounting experience, tax experience) were compared using one-way ANOVAs (comparing overall differences between means). No significant differences were detected among treatments.[9] Significant differences between cell-groups of participants on categorical demographic variables (gender, CPA certification [yes or no], current employer [Big-5 or non-Big-5], and IRS audit experience [yes or no]) were compared using chi-square tests. No significant differences between cell-groups were found.[10] Hence, each cell-group of participants appear to be representative of the sample.

Table 1. Demographic Information ($N = 153$).

GENDER:	MALE	FEMALE	TOTAL	AGE		
	87 (58%)	66 (42%)	153 (100%)	29.65 Mean (7.53) S.D.		
EDUCATION LEVEL: (highest degree)	BACHELOR	MASTER	LAW*	TOTAL		
	42 (27%)	101 (66%)	10 (7%)	153 (100%)		
MAJOR FIELD OF STUDY:	TAX	OTHER ACCTG	OTHER BUSINESS	TOTAL		
	132 (86%)	15 (10%)	6 (4%)	153 (100%)		
EXPERIENCE:	CPA	NON-CPA	TOTAL	NO. OF YRS CPA	ALL ACCT	TAX
	102 (67%)	51 (33%)	153 (100%)	4.50 Mean (6.09) Std.Dev. (CPAs only; $n = 102$)	6.41 Mean (6.68) S.D. (in years)	76.1% Mean (24.3%) S.D. (% of acct exp.)
CURRENT: EMPLOYER	LOCAL FIRM	REG'L FIRM	NATIONAL FIRM	BIG-6 FIRM	LAW FIRM	TOTAL
	73 (48%)	5 (3%)	2 (1%)	62 (41%)	11 (7%)	153 (100%)
HIGHEST POSITION:	STAFF	SENIOR	MANAGER	PARTNER	TOTAL	
	69 (45%)	38 (25%)	34 (22%)	12 (8%)	153 (100%)	
IRS AUDIT EXPERIENCE:	SOME	NONE	TOTAL	NO. OF AUDITS	SUCCESS RATE	
	77 (50%)	76 (50%)	153 (100%)	16.24 Mean (28.01) S.D. (CPAs w/IRS audit exp; $n = 77$)	78.2% Mean (26.1%) S.D. (CPAs w/IRS audit exp; $n = 77$)	
CLIENTELE: AVERAGE AGI	AGI ≤ $50K	$50K ≤ AGI ≤ $100K	$100K ≤ AGI ≤ $200K	AGI ≥ $200K	TOTAL	
	3 (2%)	13 (9%)	64 (42%)	73 (47%)	153 (100%)	
SOURCE OF INCOME	WAGES	SMALL BUSINESS	INVEST-MENTS	TOTAL		
	48 (31%)	73 (48%)	32 (21%)	153 (100%)		

* Only one subject had an advanced law degree (LLM in tax).

Test Instrument

The test instrument was used to simulate time-budget and deadline pressures that tax professionals encounter during tax season. The instrument was administered to participants on their personal computers. A two-page document containing general information about the computer program and instructions on how to execute it was distributed with each computer diskette.

Participants were given a hypothetical tax case containing four parts: (1) background material about the client and the tax issue; (2) a summary of relevant case law; (3) client information directly related to the tax issue; and (4) questions concerning participants' recommendation to the client. The test instrument also contained demographic items, manipulation check items, and a reduced version of Johnson's (1993) measure of tax-preparer aggressiveness.

Background and Case Material

The background information let the participants know that they were to resolve a tax issue for a hypothetical client by classifying him as either a dealer or investor in real estate. The real estate classification issue was chosen because it is a "gray" area of the law that is not specifically addressed in the Internal Revenue Code or Regulations. The courts have not developed specific guidelines to classify taxpayers as dealers or investors in real estate (*Los Angeles Extension Co. v. United States*, 1963; Zinicola, 1981). Generally, neither investor nor dealer classification is inherently a pro-client (aggressive) position because many factors can affect which classification is more beneficial for the client. However, in the experiment, the investor classification was the pro-client position because the client's capital loss carryovers would offset his capital gain from real estate sales.

Certain client characteristics were explicitly included in the background material to hold constant relevant information and help participants become more familiar with the client. The client was characterized as "average size" because prior research has shown that preparers are more aggressive for clients who generate large revenues for the firm (McGill, 1988; Reckers et al., 1991). The client was also represented as risk-neutral (i.e. the client had not been audited in the last 10 years, and exhibited no more than the usual aversion to an IRS audit) because studies have shown clients' risk preferences positively influence tax preparers' aggressiveness (Hite & McGill, 1992; Schisler, 1994). The background information also informed participants that they made a year-end projection for their client's tax liability based on the conservative (pro-IRS) position of being a dealer in real estate, and this information was communicated to the client.

After reading the background information, participants read a summary of case law relevant to the dealer/investor classification. This included a list of six

factors that various courts have used to make real estate classification decisions (similar to the factors used in Pei et al., 1990, 1992). This list was followed by six client facts (presented one pair per computer screen) directly related to and presented in the same order as the six classification factors. Each pair contained a pro-dealer and a pro investor fact.[11]

Dependent Variables (Client Recommendation)
Questions about the aggressiveness of tax preparers' client recommendations immediately followed the client's tax issue information. The dependent variable was the average of responses to three bipolar items (on a scale of 1 to 7): (1) What position (DEALER or INVESTOR) would you recommend to the client [1 = very strongly recommend filing as a DEALER, 7 = very strongly recommend filing as an INVESTOR]; (2) What is the likelihood that you would recommend the DEALER position to your client [1 = 0%, 7 = 100%; this item is reverse-scored to be consistent with questions 1 and 3]; (3) What is the likelihood that you would recommend the INVESTOR position to your client [1 = 0%, 7 = 100%]; The items were averaged to increase the reliability of dependent variable and to provide finer differentiation than one-item measures as the unreliability (randomness) of single-item measures is averaged out when multiple items are summed or averaged (Nunnally, 1978, p. 67). The Chronbach's alpha of the three-item dependent variable is 0.81, which is well above the minimum acceptable level of reliability of 0.70 (Nunnally, 1978, p. 245).

Experimental Manipulations

Red-flag item. Participants were exposed to a red-flag item as part of the general information about the client's tax return. The no-red-flag-item treatment stated that the client's tax return did not contain any unusual items. In contrast, the red-flag-item treatment stated that the client had a substantial home office deduction. The home-office was made unequivocally deductible to ensure that questions concerning the legality of the home-office deduction did not confound the red-flag item manipulation. Also, it is unrelated to the tax issue being analyzed so that participants in the red-flag item treatment received the same information about the tax issue as the no red-flag-item participants.

Unfavorable projection error (UPE). Tax payment status, a possible confounding variable, was held constant for UPE and no-UPE treatments, as prior studies (Duncan et al., 1989; Schisler, 1994) found payment status, payment or refund due, can cause tax preparers to think of clients' tax situations in terms of losses and gains. Since there is some support for a payment status effect, the

UPE manipulation controlled payment status by stating that a large tax payment would be due if the client was classified as a dealer in real estate (the position taken in the year-end projection) and no payment or refund would be due if the client was classified as an investor.

In the no-UPE treatment, participants were told that the year-end projection was accurate and that the large payment due was *planned.* This established a gain frame of reference, since taking the aggressive (risky) position would result in the client receiving an unexpected reduction in tax liability (an unexpected reduction of the payment due to zero). The UPE treatment, on the other hand, reflected a loss manipulation. Participants were told that they made an error on the year-end projection which caused their client to be in an *unplanned* payment due situation. They were then told that they could eliminate the effect of the UPE by claiming the more aggressive investor position, rather than the conservative dealer position.

Time pressure. The time pressure manipulation included two components. One component was included in the instrument's instructions and the other was built into the computer program. The low time-pressure participants were told to read all information carefully and take as much time as needed to complete the task. The high time-pressure participants were told that they had a very strict time budget of only four minutes to complete the task.[12] Although participants could take more than four minutes, the time budget was designed to encourage quick work. High time-pressure participants were further told that a great deal of time had already been expended on this client, giving them a "real-world" motivation to stay within the time budget. Moreover, high time-pressure participants learned that budgeted time and elapsed time would be shown continuously on the computer screen, and that a pop-up message showing budgeted time remaining would appear at regular intervals (to further induce time pressure). Elapsed time was also recorded for low time-pressure participants, but they were unaware of the timer as it was not visible.

Post-treatment Manipulation Check Items

The manipulation check items for time pressure immediately followed the timed portion of the experiment to determine if participants felt they were under time pressure while analyzing the tax issue.[13] The three time pressure items were summed to increase the reliability of the measures (Nunnally, 1978, p. 67).[14] Manipulation check items for the red-flag item[15] and UPE[16] were also included to determine if participants detected the presence of these variables.

RESULTS

Manipulation Check Items

Aggressiveness of Dealer/Investor Position

Two manipulation checks (each consisting of two manipulation check items) were used to determine whether participants perceived the investor position to be more aggressive than the dealer position. The first manipulation check compared the audit probability of the investor (aggressive) position to the dealer (conservative) position.[17] A one-tailed paired-samples t-test shows that participants perceived the audit probability to be significantly higher ($t = 6.83$; $p < 0.001$) for the aggressive investor position (mean = 3.85) than for the conservative dealer position (mean = 2.93). Another one-tailed paired-samples *t*-test also shows that participants perceived the judicial support to be significantly higher ($t = 5.53$; $p < 0.001$) for the dealer (conservative) position (mean = 4.72) than for the investor (aggressive) position (mean = 3.81).[18] Therefore, since the aggressiveness of a tax position is inversely proportional to the amount of support for that position and contrary to IRS interpretation (McGill, 1988), the results of the manipulation checks indicate that the positions were perceived as intended.

Red-Flag Item

Participants responded to two manipulation check items to determine if the red-flag item was perceived as intended (see note 12). A one-tailed independent-sample *t*-test on the sum of the two items (higher scores indicate a stronger perception a red-flag item) indicates that participants' perception of a red-flag item was significantly greater ($t = 5.40$; $p < 0.001$) for subjects who received the red-flag item treatment (mean = 8.38) than for subjects who did not receive the red-flag item treatment (mean = 4.28).

Unfavorable Projection Error (UPE)

Two manipulation check items were also used to determine if participants who received the UPE manipulation perceived that the tax projection underestimated the actual tax liability.[19] A one-tailed independent-sample *t*-test on the sum of the two items (lower scores indicate a stronger perception of a UPE) indicates that participants' perception of a UPE is significantly greater ($t = 11.32$; $p < 0.001$) for subjects who received the UPE treatment (mean = 3.32) than for subjects who did not receive the UPE treatment (mean = 7.56).

Time Pressure

Three manipulation check items were used to determine the effectiveness of the time-pressure manipulation.[20] A one-tailed independent-sample *t*-test on the sum of the three items (higher scores indicate a perception of more time pressure) indicates that participants' perception of time pressure is significantly greater ($t = 11.90$; $p < 0.001$) for subjects who received the high time-pressure treatment (mean = 15.28) than for participants who received the low time-pressure treatment (mean = 7.22). Additionally, participants in the high time-pressure treatment adhered to the time budget (mean = 4.14 minutes compared to a time budget of four minutes), while low time-pressure participants (who had no time budget) took significantly longer ($t = 6.62$; $p < 0.001$) to complete the instrument (mean = 6.05 minutes).

Hypothesis Tests

The fully-crossed ANOVA model shown in Table 2 is significant ($F = 6.10$; $p < 0.001$) with an adjusted R-Squared of 0.19.[21] The ANOVA model also shows a three-way interaction among the red-flag item, UPE, and time pressure. Due to the significant three-way interaction and lower order interactions, the hypotheses were analyzed by comparing specific treatment means using one-tailed *post hoc* pairwise comparisons (see Table 3).[22] This analytic method is further supported by the power analysis in Table 2, which shows that the observed power for two of the three main effects is very low (red-flag item = 0.277; UPE = 0.881; time pressure = 0.053) while the observed power of the three-way interaction is extremely high (0.924).

Table 2. ANOVA Results of Red-Flag Item by UPE by Time Pressure.
Dependent Variable: Tax Preparer Recommendation.

Source of Variation	SS	DF	MS	*F*	*p*-value	Observed Power
Within + Residual	176.05	145	1.21			
Red-flag	2.30	1	2.30	1.90	0.171	0.277
UPE	12.13	1	12.13	9.99	0.002	0.881
Time Pressure	0.04	1	0.04	0.03	0.865	0.053
Red-flag*UPE	12.98	1	12.98	10.69	0.001	0.901
Red-flag*Time Pressure	4.38	1	4.38	3.61	0.059	0.471
UPE*Time Pressure	5.17	1	5.17	4.26	0.041	0.536
Red-flag*UPE*Time Pressure	14.18	1	14.18	11.68	0.001	0.924
Model	51.88	7	7.41	6.10	0.001	0.999
Total	227.92	152	1.50			
R-Squared	0.228					
Adjusted R-Squared	0.190					

Table 3. Cell Means for ANOVA Model.
Dependent Variable: Tax Preparer Recommendation.

	Low Time Pressure		High Time Pressure		Marginal Means
	No UPE	UPE	No UPE	UPE	
No Red-Flag Item	4.15 $n = 20$ Cell 1	3.89 $n = 18$ Cell 3	4.28 $n = 18$ Cell 5	4.50 $n = 18$ Cell 7	4.20
Red-Flag Item	3.05 $n = 20$ Cell 2	5.18 $n = 19$ Cell 4	3.72 $n = 19$ Cell 6	3.89 $n = 21$ Cell 8	3.96
Marginal Means	4.07		4.10		

Marginal Means	
No UPE	3.80
UPE	4.36

1 = most conservative recommendation.
7 = most aggressive recommendation.

H1: Test of Red-Flag Item

To support H1 – which predicts that participants will make more conservative recommendations when a red-flag item is present – participants' scores should be lower when the red-flag item is present than when it is absent for each combination of UPE and time pressure. As shown in Table 4, participants in three of the four comparisons made more conservative recommendations when the red-flag item was present, but only two of the one-tailed comparisons were significant at the 0.05 level ($p < 0.001$, $p < 0.063$, and $p < 0.043$). These comparisons indicate that the influence of a red-flag item on tax preparers' aggressiveness may be dependent on omitted variables that are present in the decision-making environment. Therefore, H1 is partially supported.

H2: Test of Unfavorable Projection Error (UPE)

To support H2 – which predicts that participants will make more aggressive recommendations when a UPE is present – participants' scores should be higher when the UPE is present than when it is absent for each combination of the

Table 4. H1: Post Hoc Multiple Pairwise Comparisons of Cell Means Comparison of Red-Flag Treatments to No-Red-Flag Treatments.
Dependent Variable: Tax Preparer Recommendation.

Description	Mean Difference	Significance
No UPE, Low Time Pressure	−1.10	$p < 0.001$
UPE, Low Time Pressure	+1.29	$p < 0.001$ wrong direction
No UPE, High Time Pressure	−0.56	$p < 0.063$
UPE, High Time Pressure	−0.61	$p < 0.043$

red-flag item and time pressure. As indicated in Table 5, participants made more aggressive recommendations when UPE was present only when the red-flag item was also present under low time pressure ($p < 0.001$). Therefore, H2 is partially supported.

H3: Time Pressure

To support H3 – which predicts that participants will make more aggressive recommendations under high time pressure than under low time pressure – participants' scores should be higher for each combination of the red-flag item and UPE. As reflected in Table 6, participants made more aggressive recommendations when either the red-flag item or the UPE was present ($p < 0.030$, $p < 0.049$). Hence, H3 is partially supported.

Table 5. H2: Post Hoc Multiple Pairwise Comparisons of Cell Means Comparison of UPE Treatments to No-UPE Treatments.
Dependent Variable: Tax Preparer Recommendation.

Description	Mean Difference	Significance
No Red-Flag Item, Low Time Pressure	−0.26	$p < 0.234$
Red-Flag Item, Low Time Pressure	+2.13	$p < 0.001$
No Red-Flag Item, High Time Pressure	+0.22	$p < 0.523$
Red-Flag Item, High Time Pressure	+0.17	$p < 0.564$

Table 6. Test of H3: Post Hoc Multiple Pairwise Comparisons of Cell Means Comparison of High Time-Pressure Treatments to Low Time-Pressure Treatments.
Dependent Variable: Tax Preparer Recommendation.

Description	Mean Difference	Significance
No Red-Flag Item, No UPE	+0.13	$p < 0.611$
Red-Flag Item, No UPE	+0.67	$p < 0.030$
No Red-Flag Item, UPE	+0.61	$p < 0.049$
Red-Flag Item, UPE	−1.29	$p < 0.001$ wrong direction

Surprisingly, when the red-flag item and the UPE were present, time pressure resulted in *less* aggressive reporting ($p < 0.001$), which is in opposite of the predicted direction. An examination of Fig. 1 shows that extremely aggressive recommendations made by participants under low time pressure with both the red-flag item and UPE present created an unexpected three-way interaction (see Table 3). It appears that participants under low time either ignored the red-flag item when the UPE was present, or perhaps without time constraints, the UPE caused them to find a way to use the red-flag item to support the aggressive position. However, one cannot rule out the possibility that one or more of the manipulations were flawed.

SUMMARY AND CONCLUSIONS

Discussion of Results

Test of H1: Red-Fag Item

The tests of H1 showed a link between the presence (or absence) of an item on a tax return believed to attract IRS scrutiny (a legitimate home office deduction) and a preparer's recommendation when resolving an ambiguous tax issue. This is not surprising as related research on audit probability only found audit probability to have an effect when it was explicitly stated as either high (McGill, 1988) or 50% (Kaplan et al., 1988), but not when audit probability was stated to be 25% (Duncan et al., 1989; LaRue & Reckers, 1989).

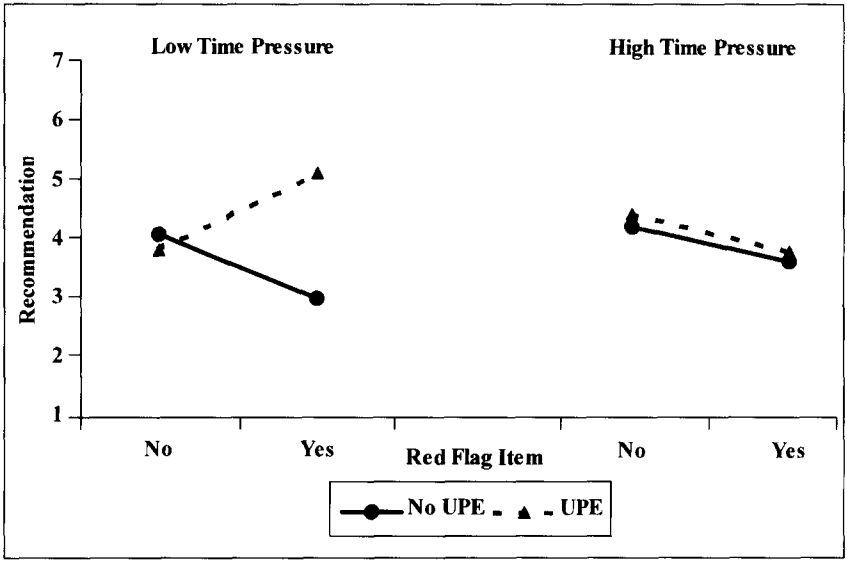


Fig. 1. Graph of Cell Means for ANOVA Table Dependent Variable: Tax Preparer Recommendation.

Also, while this study does not specifically manipulate audit probability, the interaction between the red-flag item and UPE is similar to the finding in McGill (1988) and Kaplan et al. (1992) in which other variables interacted with audit probability. Although client importance in McGill and tax experience in Kaplan et al. only affected the magnitude of the effect of audit probability on tax preparers' recommendations, unfavorable projection error combined with the red-flag item affected the direction of the effect of the red-flag item under low time pressure (participants made more *aggressive* recommendations when the red-flag item was present). These two studies support the findings in the current study that tax preparers may make more conservative recommendations when perceived IRS scrutiny is high, but other factors may affect this relationship.

Test of H2: Unfavorable Projection Error (UPE)

The second hypothesis predicted that underestimating a client's tax liability by making a mistake on his/her year-end projection would cause a tax preparer to make an aggressive recommendation on an ambiguous issue to compensate for the mistake. Although H2 was only partially supported (one of the four comparisons in Table 5 were significant), UPE was significant in one of the low time-pressure comparisons, but in neither high time-pressure comparison.

This may indicate that the framing effect was stronger under low time pressure (when the red-flag item was present) than under high time pressure, which is consistent with Svenson and Benson (1993, Experiment 1) who found that time pressure reduced the effects of framing on participants decisions. The nature of the observed relationships among the variables will be discussed further when the three-way interaction between the red-flag item, UPE, and time pressure is addressed.

Test of H3: Time Pressure

Although H3 was only supported at the 0.05 significance level in two of the four comparisons, high time-pressure participants' recommendations were significantly different (at the 0.05 level) than their low time-pressure counterparts in three of the four comparisons (see Table 6). The only non-significant comparison involved subjects who received only issue-relevant information (no-red-flag item and no UPE). However, subjects consistently used less time under high time pressure than under low time pressure as indicated by the comparisons of treatment means of a fully-crossed ANOVA with time used as the dependent variable (see Tables 7, 8 and 9).

Figure 1 illustrates the effect of time pressure on tax preparers' recommendations. Under low time pressure, participants spent significantly more time when both the red-flag item and UPE were present (see Table 10), but high time-pressure participants who received both red-flag and UPE treatments did not spend more time than other high time-pressure participants (see Table 11).

Table 7. ANOVA Results of Red-Flag Item by UPE by Time Pressure. Dependent Variable: Time Used.

Source of Variation	SS	DF	MS	F	*p*-value	Observed Power
Within + Residual	439.68	145	3.03			
Red-flag	9.14	1	9.14	3.01	0.085	0.407
UPE	5.47	1	5.47	1.80	0.181	0.266
Time Pressure	138.29	1	138.29	45.60	0.001	1.000
Red-flag*UPE	0.94	1	0.94	0.31	0.879	0.053
Red-flag*Time Pressure	11.73	1	11.73	3.87	0.051	0.498
UPE*Time Pressure	4.68	1	4.68	1.54	0.216	0.235
Red-flag*UPE*Time Pressure	7.59	1	7.59	2.50	0.116	0.349
Model	177.84	7	25.41	8.38	0.001	1.000
Total	617.52	152				
R-Squared	0.288					
Adjusted R-Squared	0.254					

Table 8. Cell Means of Time Used.
Dependent Variable: Time Used.
(Time Used Reported in Minutes)

	Low Time Pressure		High Time Pressure		Marginal Means
	No UPE	UPE	No UPE	UPE	
No Red-Flag Item	5.41 $n = 20$ Cell 1	5.65 $n = 18$ Cell 3	3.96 $n = 18$ Cell 5	4.39 $n = 18$ Cell 7	4.85
Red-Flag Item	5.96 $n = 20$ Cell 2	7.18 $n = 19$ Cell 4	4.30 $n = 19$ Cell 6	3.93 $n = 21$ Cell 8	5.34
Marginal Means	6.05		4.15		

Marginal Means	
No UPE	4.91
UPE	5.29

Table 9. Test of Time Used: Post Hoc Multiple Pairwise Comparisons of Cell Means Comparison of High Time-Pressure Treatments to Low Time-Pressure Treatments.

Dependent Variable: Time Used.
(Time Used Reported in Minutes)

Description	Mean Difference	Significance
No Red-Flag Item, No UPE	−1.45	$p < 0.006$
Red-Flag Item Present, No UPE	−1.66	$p < 0.002$
No Red-Flag Item, UPE	−1.57	$p < 0.016$
Red-Flag Item Present, UPE	−3.25	$p < 0.001$

Table 10. Test of Time Used: *Post Hoc* Multiple Pairwise Comparisons of Cell Means Comparison of Red-Flag Treatments by UPE Treatments under Low Time-Pressure Against other Low Time-Pressure Treatments.
Dependent Variable: Time Used.
(Time Used Reported in Minutes)

Description	Mean Difference	Significance
Red-Flag Item, UPE No Red-Flag Item, No UPE	+1.77	$p < 0.001$
Red-Flag Item, UPE Red-Flag Item, No UPE	+1.22	$p < 0.016$
Red-Flag Item, UPE No Red-Flag Item, UPE	+1.53	$p < 0.004$

It appears that high time-pressure participants encountered a pseudo-ceiling effect that was created by the instructions. That is, participants in the high time pressure condition were likely to use the entire four-minute budget because that was just enough time to quickly and thoroughly read the information, but they were not likely to exceed the budget because the instructions stated that they had a very strict time budget of only four minutes to complete the task. The complex relationships described above will be further addressed by specifically analyzing the three-way interaction between the red-flag item, UPE, and time pressure.

Interaction between Red-Flag Item, UPE, and Time Pressure

In this study, only main effects were hypothesized for the three independent variables, but a three-way interaction was indicated by the results. As indicated by Table 12 (Panel A), the red-flag item and UPE interacted under low time pressure ($p < 0.001$), but not under high time pressure (Table 12, Panel B; $p < 0.923$).

Under low time pressure, the red-flag item had no significant effect when the UPE was present. An analysis of the red-flag manipulation check items reveal that participants who received both the red-flag item and UPE under low

Table 11. Test of Time Used: Post Hoc Multiple Pairwise Comparisons of Cell Means Comparison of Red-Flag Treatments by UPE Treatments under High Time-Pressure Against other High Time-Pressure Treatments.
Dependent Variable: Time Used (Time Used Reported in Minutes).

Description	Mean Difference	Significance
Red-Flag Item, UPE No Red-Flag Item, No UPE	−0.03	$p < 0.724$
Red-Flag Item, UPE Red-Flag Item, No UPE	−0.37	$p < 0.249$
Red-Flag Item, UPE No Red-Flag Item, UPE	−0.46	$p < 0.202$

Table 12. ANOVA Results of Red-Flag Item by UPE.
Dependent Variable: Tax Preparer Recommendation

Panel A: Low Time Pressure

Source of Variation	SS	DF	MS	*F*	*p*-value
Within + Residual	76.25	73	1.04		
Red-flag	0.17	1	0.17	0.16	0.690
UPE	16.70	1	16.70	15.98	0.001
Red-flag*UPE	27.36	1	27.36	26.19	0.001
Model	44.73	3	14.91	14.27	0.001
Total	120.98	76	1.59		
R-Squared	0.370				
Adjusted R-Squared	0.344				

Panel B: High Time Pressure

Source of Variation	SS	DF	MS	*F*	*p*-value
Within + Residual	99.80	72	1.39		
Red-flag	6.47	1	6.47	4.67	0.034
UPE	0.73	1	0.73	0.52	0.471
Red-flag*UPE	0.01	1	0.01	0.01	0.923
Model	7.12	3	2.37	1.71	0.172
Total	106.92	75	1.43		
R-Squared	0.067				
Adjusted R-Squared	0.028				

time pressure (Cell 4 of Table 3) *did not* report a significantly higher presence of a red-flag item than did participants in three of the four no red-flag item conditions (perception of the red-flag item when the UPE was present was only significantly greater than when neither red-flag item nor UPE was present [Cell 1]). However, the opposite results occurred under high time pressure. That is, participants who received the red-flag item and the UPE (Cell 8 of Table 3) *did* report a significantly higher presence of a red-flag item than participants in three of the four no red-flag item conditions (perception of the red-flag item when the UPE was present was *not* significantly greater than when neither red-flag item nor UPE was present [Cell 5]). Additionally, a cell-by-cell analysis of the UPE manipulation check items revealed that time pressure did *not* affect the perception of the UPE manipulation (all differences between cell means were as expected). These results seem to indicate that participants under low time pressure either chose to focus on the UPE instead of the red-flag item or did not perceive the home office deduction to be a red-flag item. It is possible that these results were due to a weak manipulation of the red-flag item, but, this does not seem likely as perception of the red-flag item in the red-flag item conditions was significantly greater than the no red-flag item conditions in 12 of the 16 possible comparisons.

Analysis of three different dependent variables indicate that time pressure had an effect on the decision-making environment: the aggressiveness of participants' recommendations, participants' perception of the red-flag item, and the amount of time used. Even if these differences were due to a weak manipulation, it would seem that the effectiveness of the manipulation was affected by the level of time pressure as well as the presence (or absence) of other variables. Therefore, this unexpected interaction indicates that time pressure does affect accountant's judgments, but the nature of the relationship is difficult to predict.

Contributions

This study makes two contributions related to the research variables. First, this study contributes to the audit probability literature by showing in an experimental setting that tax professionals generally made less aggressive recommendations in the presence of a frequently encountered item believed to attract IRS scrutiny (a home office deduction). The moderate success of this manipulation advances audit probability research as this study uses a "real world" indicator of increased IRS attention instead of the artificial manipulation previously used – explicitly stated audit probability (e.g. Kaplan et al., 1988; McGill, 1988; Duncan et al., 1989; LaRue & Reckers, 1989).

Second, this study indicates that time pressure may affect tax preparers' recommendations on ambiguous tax issues. Although the time-pressure hypothesis was only partially supported, time pressure did significantly affect preparers' recommendations in three of the four comparisons (at the 0.05 level) when compared to the recommendations of low time-pressure participants who received the same treatments. These comparisons indicate that participants' responses under low time pressure were different from high time-pressure participants' responses (see Fig. 1). Therefore, additional research is needed to gain a better understanding of how time pressure affects tax preparers' decisions.

Limitations

In this study participants analyzed only one tax issue – classifying a taxpayer as a dealer or an investor in real estate. Therefore, the results obtained may not be generalizable to other tax issues. Future studies on time pressure should incorporate different types of tax issues (e.g. income items, deductions). Also, it is not known if participants' responses to a hypothetical tax scenario with no "real world" consequences are representative of their behavior under actual working conditions. However, the use of actual case law and a realistic client situation coupled with a large and diverse sample of tax professionals minimize the effect of this limitation.

A final limitation is that this study allowed tax preparers to make recommendations to a client based on an incremental scale. This was done to increase the variance of the responses even though Sommerfeld et al. (1989) stated that in actual tax practice, preparers are usually faced with a dichotomous choice of either recommending taking or not taking a deduction. Future studies should be conducted using a dichotomous dependent variable – recommend, do not recommend – so that even participants who are relatively neutral must make a choice one way or the other.

NOTES

1. Avoiding a preparer penalty under §6694 only requires the position to have at least a one-in-three possibility of prevailing on its merits (§10.34(a)(1) of the Treasury Department's Circular 230; IRS, 1994). Therefore, an ambiguous item would meet these criteria because it would have approximately a 50% likelihood of being upheld. Moreover, the taxpayer penalty under §6662 would also likely not apply because Reg. §1.6662-4(d)(2) states that the substantial authority requirement for §6662 is "less stringent than the more likely than not standard (. . . greater than 50% likelihood of the position being upheld)."

2. The audit probability hypothesis was supported when the high audit probability treatment was 50% (Kaplan et al., 1988), but not when it was only 25% (Duncan et al., 1989; LaRue & Reckers, 1989). Also, McGill (1988) supported the hypothesis when participants were informed that audit probability was "high."

3. If a preparer does not know the client, he or she is likely unaware of the total fees generated by that client for other work (tax, write-up, audit, and consulting for the taxpayer and related entities). Moreover, a client's billing statements are not usually included in a client file.

4. Thorndike's (1913) Law of Effect is discussed in most management principles textbooks (e.g. Bateman & Snell, 1999).

5. For example, 89% of tax accountants surveyed in Smith and Hutton (1995) report under-reporting chargeable time at least occasionally.

6. See Svenson and Maule (1993) for a review of time-pressure literature.

7. This information was self-reported by the auditor participants and supported by their empirical results.

8. One hundred sixty-seven instruments were returned. Six of the returned instruments were not usable because they were not completely filled-out. Another seven participants were excluded because they did not have sufficient tax experience (less than 40% of their time was devoted to tax work). One additional participant was excluded because he/she took three hours to complete the task, which indicated that he/she might have violated the instructions by consulting colleagues or reference materials.

9. The results of the F tests of the one-way ANOVAs are as follows: age $F = 0.612$, $p < 0.745$; years of accounting experience $F = 0.383$, $p < 0.911$; tax experience as a percentage of total accounting experience $F = 0.604$, $p < 0.752$.

10. The results of the chi-square tests are as follows: gender $\chi_7^2 = 3.16$, $p < 0.874$; CPA certification $\chi_7^2 = 4.49$, $p < 0.722$; current employer $\chi_7^2 = 3.13$, $p < 0.873$; IRS audit experience $\chi_7^2 = 4.21$, $p < 0.755$.

11. Participants were randomly assigned to one of two information presentation orders. Eighty participants received fact pairs with the pro-dealer information first and 73 participants' fact pairs listed the pro-investor information first. The difference between means of the dependent variable of these two groups was not statistically significant ($p < 0.49$). This was done to control for information presentation order as found in Pei et al. (1990, 1992).

12. A pilot test of tax professionals was used to determine the amount of time needed to complete the instrument if the information was read quickly yet carefully one time through.

13. The following three items were used to measure participants' perceived time pressure. How much time pressure did you feel when performing this task (1 = Very LITTLE time pressure; 7 = Very MUCH time pressure)? How much impact did the time budget have on the amount of time that you took to complete the task (1 = Very LITTLE Impact; 7 = Very MUCH impact)? How much time pressure did you feel to finish the task within the time budget (1 = Very LITTLE time pressure; 7 = Very MUCH time pressure)?

14. The inter-item correlations for the three time-pressure variables ranged from 0.85 to 0.87. Also, Cronbach's Alpha for the three items was 0.95.

15. The following two items were used to measure participants' perception of the red-flag item. Did you notice any item on the tax return (NOT the dealer/investor issue) that is believed to attract IRS attention (1 = NO; 7 = YES). Was there any item on the client's tax return (NOT the dealer/investor issue) that you feel would be a "RED FLAG"

to the IRS (1 = NO; 7 = YES)? The inter-item correlation for the two red-flag item variables was 0.78 and the Cronbach's Alpha was 0.88.

16. The following two items were used to measure participants' perception of the unfavorable projection error. How accurate was the year-end tax projection with respect to the actual tax liability, assuming that the client files as a dealer (1 = Tax projection UNDERSTATED actual liability; 4 = Tax projection very accurate; 7 = Tax projection OVERSTATED actual liability). The client's year-end tax projection was based on being a dealer in real estate. If the client files as a dealer, how would his ACTUAL tax payment status (payment or refund due) change with respect to his PROJECTED tax payment status? (1 = Larger PAYMENT Due; 4 = No change in payment status; 7 = Larger REFUND Due)? The inter-item correlation for the two UPE variables was 0.73 and the Cronbach's Alpha was 0.84.

17. Audit probability of the investor and dealer positions were measured on a 7-point scale (1 = very unlikely; 7 = very likely) by the following two questions. If the client files as an INVESTOR in real estate, what is the likelihood that the tax return will be audited? If the client files as a DEALER in real estate, what is the likelihood that the tax return will be audited?

18. Judicial support for the dealer and investor positions were measured on a 7-point scale (1 = very unlikely; 7 = very likely) by the following two questions. What is the likelihood that the DEALER position will be JUDICIALLY supported? What is the likelihood that the INVESTOR position will be JUDICIALLY supported?

19. The two manipulation check items (see note 14) are measured on a 7-point scale such that a response of "4" indicates that the tax projection was accurate (corresponding to the no UPE manipulation), a response less than four indicates that the tax projection understated the actual tax liability (corresponding to the UPE manipulation), and a response greater than four indicates that the tax projection overestimated the actual tax liability (corresponding to neither manipulation). The range of possible scores is from 2–14, but respondents should not report scores above eight because they do not correspond to either manipulation.

20. The range of possible scores is 3–21 (see note 11 for the manipulation check items).

21. None of the following variables were significantly related to the dependent variable, and therefore not included as covariates: age, gender, aggressiveness, current position, CPA status (yes or no), Big-5 experience, tax experience (linear and non-linear transformations), IRS audit experience, IRS success, client size, client type, and information presentation order.

22. The Least Significant Difference method was used to reduce the chance of a Type I error when performing multiple post hoc pairwise comparisons.

ACKNOWLEDGMENTS

The authors wish to thank the editor, associate editor, and anonymous reviewers for their insightful comments. Additionally, the authors acknowledge the comments and suggestions provided by participants at the 1999 AAA National Meeting (San Diego, CA) and 1999 22nd Annual Conference of the European Accounting Association (Bordeaux, France).

REFERENCES

American Institute of Certified Public Accountants (1997) *AICPA Professional Standards*, Vol. 2. Chicago: Commerce Clearing House.

Anderson, U., Koonce, L., & Marchant, G. (1994). The effects of source-competence information and its timing on auditors' performance of analytical procedures. *Auditing: A Journal of Practice & Theory, 13*, 137–148.

Bateman, T. S., & Snell, S. A. (1999). *Management: Building Competitive Advantage*. Boston: Irwin McGraw-Hill.

Brown, C., & Solomon, I. (1992). Auditors judgments/decisions under time-pressure: An agenda for research and an illustration. In: R. Srivastava (Ed.), *Auditing Symposium XI: Proceedings of the 1992 Deloitte and Touche-University of Kansas Symposium on Auditing Problems*. Lawrence, KS: University of Kansas.

Chaiken, S. (1980). Heuristic vs. systematic information processing and the use of source vs. message cues in persuasion. *Journal of Personality and Social Psychology, 39*, 752–766.

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic information processing within and beyond the persuasion context. In: J. S. Uleman & J. A. Bargh (Eds), *Unintended Thought* (pp. 212–252). New York: Guilford Press.

Choo, F., & Firth, M. (1993). The effect of time-pressure on auditors' configural information processing. Unpublished manuscript.

Christensen-Szalanski, J. J. (1980). A further examination of the selection of problem-solving strategies: The effects of deadlines and analytic aptitudes. *Organizational Behavior and Human Performance, 25*, 107–122.

Douglas, S. (1979). An examination of underreporting of hours by accountants within an expanded expectancy theory framework. Unpublished doctoral dissertation, University of Southern California.

Duncan, W. A., LaRue, D., & Reckers, P. M. J. (1989). An empirical examination of the influence of selected economic and non-economic variables on decision-making by tax practitioners. *Advances in Taxation, 2*, 91–106.

Flanagan, W. G. (Ed.) (1981). IRS vs. taxpayers: tales of the tape. *Forbes, 128*(25), 162–165.

Hite, P. A., & McGill, G. A. (1992). An examination of taxpayer preference for aggressive tax advice. *National Tax Journal, 45*, 389–403.

Internal Revenue Code of 1986 (1995). St. Paul, MN: West Publishing Company.

Internal Revenue Service (1994). Circular 230: Regulations Governing the Practice of Attorneys, Certified Public Accountants, Enrolled Agents, Enrolled Actuaries, and Appraisers Before the Internal Revenue Service.

Jenkins, B., & Schuch, B. (1997). How to avoid a tax audit. Take It Personally, April 8, transcript no. 977040806FN-108, Cable News Network Financial.

Johnson, L. M. (1993). An empirical investigation of the effects of advocacy on preparers' evaluations of judicial evidence. *Journal of the American Taxation Association, 15*(1), 1–22.

Jones, T. M. (1991). Ethical decision making by individuals in organizations: An issue-contingent model. *Academy of Management Review, 16*(2), 366–395.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica, 47*(2), 263–291.

Kaplan, S. E., Reckers, P. M. J., West, S. G., & Boyd., J. H. (1988). An examination of tax reporting recommendations of professional tax preparers. *Journal of Economic Psychology, 9*, 427–443.

Kastantin, J. T. (1988). *Professional Accounting Practice Management.* New York: Quorum Books.

Kelly, T., & Margheim, L. (1990). The impact of time budget pressure, personality, and leadership variables on dysfunctional auditor behavior. *Auditing: A Journal of Practice and Theory, 9,* 21–42.

LaRue, D., & Reckers, P. M. J. (1989). An empirical examination of the influence of selected factors on professional tax preparers' decision processes. *Advances in Accounting, 7,* 37–50.

Lightner, S. M., Adams, S. J., & Lightner, K. M. (1982). The influence of situational, ethical, and expectancy theory variables on accountants' underreporting behavior. *Auditing: A Journal of Practice and Theory, 2,* 1–12.

Lohse, D. (1994, February 25). Bulletproof your income-tax return to avoid becoming an IRS target. *The Wall Street Journal,* pp. C1, C18.

Los Angeles Extension Co. v. United States (1963). United States Court of Appeals Ninth Circuit. 63-1 U.S. Tax Case (CCH) ¶9365.

McGill, G. (1988). The CPA's role in income tax compliance: an empirical study of variability in recommending aggressive tax positions. Unpublished doctoral dissertation.

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill Publishing Company.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 534–552.

Pei, B. K. W., Reckers, P. M. J., & Wyndelts, R. W. (1990). The influence of information presentation order on practitioner judgment. *Journal of Economic Psychology, 11,* 119–146.

Pei, B. K. W., Reckers, P. M. J., & Wyndelts, R. W. (1992). Tax practitioners belief revision: The effects of information presentation sequence, client preference, and domain experience. *Decision Sciences, 23,* 175–199.

Raabe, W. A, Whittenburg, G. E., & Bost, J. C. (1997). *West's Federal Tax Research* (4th ed.). St. Paul, MN: West Publishing Company.

Raby, W. L. (1974). *Tax Practice Management.* New York: American Institute of Certified Public Accountants.

Rachlin, N. S. (1983). *Eleven Steps to Building a Profitable Accounting Practice.* New York: McGraw-Hill.

Ratneshwar, S., & Chaiken, S. (1991). Comprehension's role in persuasion: the case of its moderating effect on the persuasive impact of source cues. *Journal of Consumer Research, 18,* 52–63.

Reckers, P. M. J., Sanders, D. L., & Wyndelts, R. W. (1991). An empirical investigation of factors influencing tax practitioner compliance. *Journal of the American Taxation Association, 13*(2), 30–46.

Rothstein, H. G. (1986). The effects of time pressure on judgment in multiple cue probability learning. *Organizational Behavior and Human Decision Processes, 37,* 83–92.

Schisler, D. L. (1994). An experimental examination of factors affecting tax preparers' aggressiveness – a Prospect Theory approach. *Journal of the American Taxation Association, 16,* 124–142.

Singhapakdi, A., Vitall S. J., & Franke, G. R. (1999). Antecedents, consequences, and mediating effects of perceived moral intensity and personal moral philosophies. *Journal of the Academy of Marketing Science, 27,* 19–35.

Smith, R. W., & Hutton, M. R. (1995). Underreporting time: An analysis of current tax practice. *Journal of Applied Business Research, 11,* 39–45.

Sommerfeld, R. (1989). *Tax Research Techniques.* New York: American Institute of Certified Public Accountants.

Spilker, B. C. (1995). The effects of time pressure and knowledge on key work selection behavior in tax research. *The Accounting Review, 70*, 49–70.

Spilker, B. C., & Prawitt, D. F. (1997). Adaptive responses to time pressure: the effects of experience on tax information search behavior. *Behavioral Research in Accounting, 9*, 172–198.

Svenson, O., & Benson, L., III. (1993). Framing and Time Pressure in Decision Making. In: O. Svenson & A. J. Maule (Eds), *Time Pressure and Stress in Human Judgment and Decision Making*. New York: Plenum Press.

Svenson, O., & Edland, A. (1987). Change of preferences under time-pressure: Choices and judgments. *Scandinavian Journal of Psychology, 28*, 322–330.

Svenson, O., & Maule, A. J. (1993). *Time Pressure and Stress in Human Judgment and Decision Making*. New York. Plenum Press.

Thorndike, E. L. (1913). *The Psychology of Learning*. New York: Teachers College.

Wartzman, R. (1993, February 24). Don't wave a red flag at the IRS. *The Wall Street Journal*, pp. C1, C18.

Wiltsee, J. L. (Ed.) (1984, April 9). Get ready: a tax audit is more likely than you think. *Business Week*, 121–122.

Wright, P. (1974). The harassed decision maker: Time pressure, distraction and the use of evidence. *Journal of Applied Psychology, 59*, 555–561.

Wright, P., & Weitz, B. (1977). Time horizon effects on product evaluation strategies. *Journal of Marketing Research, 14*, 429–443.

Zinicola, P. A. (1981). Real estate and section 1221: Business as a pattern of activity in the definition of a capital asset. *The Tax Lawyer, 35*, 225–256.

THE KEIRSEY TEMPERAMENT SORTER: INVESTIGATING THE IMPACT OF PERSONALITY TRAITS IN ACCOUNTING

Patrick Wheeler, Carol Jessup and Michele Martinez

ABSTRACT

The Keirsey Temperament Sorter (KTS) is a psychometric instrument that can be useful to researchers interested in investigating the impact of personality traits in accounting practice and education. The KTS may be used to investigate: (a) the nature of personality of accounting practitioners, faculty and students; and (b) how personality traits affect the performance of accounting practitioners, faculty and students. This paper provides KTS users with the empirical and conceptual information necessary to conduct research in these areas. The psychometric properties and underlying theory of the KTS are examined, as are its limitations. The paper also compares the KTS to the Myers-Briggs Type Indicator (MBTI), another widely used psychometric instrument. The two instruments are compared in order to determine the relative advantages of each for accounting research.

Advances in Accounting Behavioral Research, Volume 5, pages 247–277.

ISBN: 0-7623-0953-9

INTRODUCTION

Instruments for determining psychological characteristics and personality traits have been used in accounting research, accounting education and the accounting profession for several decades (Wheeler, 2001). Some of these instruments derive from personality theories, a cluster of theories comprising one of the established branches of psychology (Feist & Feist, 1998; Hergenhahn & Olson, 1999). For example, the Myers-Briggs Type Indicator (MBTI) has been used in industry and psychology for over 40 years and in accounting research for over 20 years (Myers et al., 1998; Wheeler, 2001). This instrument is based on C. G. Jung's theory of personality type.

Another more recent personality trait instrument is the Keirsey Temperament Sorter (KTS). The KTS is based on temperament theory, a personality theory that is indirectly related to Jungian personality type theory. The KTS, like the MBTI, has been used by businesses and in education and has recently begun appearing in published academic accounting research (e.g. Gul & Fong, 1993). The purpose of this paper is to examine the theoretical and practical aspects of the KTS, especially in comparison to the MBTI. Accordingly, both temperament theory and Jung's personality type theory will be discussed. The research and educational applications of the KTS will also be examined in detail.

From a research perspective several issues surround the use of the KTS as an instrument for conducting behavioral research. These issues may be categorized as: (1) the theoretical underpinning of the instrument; (2) the psychometric properties of the instrument; and (3) the areas of application in which the instrument may be used. The first category of issues is concerned with the constructs that the KTS purports to capture and measure. While the logical and philosophical integrity of a theory is important, the ability of an instrument to measure variables derived from theory is equally important, particularly from an empirical perspective, since this capability directly impacts how the theory is applied. There are numerous psychological instruments, none of which can reasonably claim to capture the whole spectrum of personality traits; therefore, the experimenter must carefully choose the instrument to match the specific needs of the experiment being conducted.

The second category, psychometric properties, is concerned with reliability and validity of the instrument. Reliability testing examines the consistency of the instrument over several dimensions; e.g. time, populations and different versions of the instrument. Validity testing is concerned with the theoretical soundness of the instrument.

The third category of issues, areas of application, examines the types of questions that can be investigated by the instrument. First, as noted above, the

KTS is not suitable for investigating all psychological or even personality research questions. For example, according to theory, the personality traits in temperament theory are predicted to be highly stable over time, if not immutable. Thus, researchers should not use the KTS to capture before-and-after treatment effects. Second, the KTS differs from Jungian-based instruments, in particular the MBTI. The KTS and the MBTI should not be applied in the same manner. Both the KTS and the MBTI purport to measure the four Jungian personality bipolar scales: Extraversion-Introversion; Sensing-Intuition; Thinking-Feeling; Judging-Perceiving, respectively.[1] However, while the MBTI is especially concerned with the four-way interaction of the bipolar scales as the basis for the personality *types*, the KTS is concerned with two-way interactions as the basis for personality *temperaments*.[2]

In this paper, these three categories of issues will be discussed, with emphasis on their significance to accounting. The KTS is an instrument with much potential for furthering accounting research. To use this instrument properly, understanding its underlying theory and its psychometric properties is necessary. Understanding the relationship between the KTS and the MBTI is also helpful, both how they overlap and differ. Once these areas are examined, the wide range of potential research topics made possible by the KTS can be explored.

The remainder of this paper will first discuss temperament theory, the personality theory underlying the KTS, followed by an examination of Jung's theory of personality types. In the context of this paper, there are two reasons for examining Jungian theory. First, although temperament theory is distinct from Jungian theory, the two are closely, if indirectly, related. This complex relationship allows the KTS, as an instrument, to use variables developed by Jungian theory as the specific personality traits it attempts to measure. Thus, both the KTS and MBTI measure the same personality-trait variables, although they interpret them and their interactions differently. The second reason for discussing Jungian theory is the need to compare the psychometric properties and experimental uses of the two instruments. Following the discussion of Jungian theory, the KTS as a psychometric instrument is examined in detail, including its validity and reliability. The KTS is compared to the MBTI in this section. Next, research opportunities using personality theory instruments are delineated, with particular attention paid to the KTS. Finally, there are concluding comments, including the limitations of using the KTS and other personality trait instruments for accounting research.

TEMPERAMENT THEORY

Temperament theory postulates four fundamental temperaments as the source of many observable human characteristics and predictable behaviors.[3] A

temperament may be defined as a cluster of distinct and interrelated personality characteristics or traits. While an individual will have a mix of some or all of the four temperaments, one temperament will clearly dominate the personality according to temperament theory (Keirsey, 1998; Berens, 1996, 1998). This mixture of the temperaments, including the dominant temperament, is theorized to be inherited. Therefore, the temperament structure (i.e. mixture and dominance) is a highly stable aspect of the personality. Although the temperament undergoes development as the individual matures, temperament theory postulates that the basic structure stays relatively unchanged.

Temperament theory has a history dating back to at least Hippocrates' theory of the four humors in the 5th century BC. A review of temperament theory indicates that the theory has influenced a wide range of disciplines throughout the centuries, including modern philosophy, psychology and medical science (Roback, 1973). Temperament theory influenced Jung's formulation of his particular personality type theory (Jung, 1971; Keirsey, 1998). This link is especially relevant because this indirect relationship allows the KTS to utilize personality variables defined in Jungian psychology, although they are not strictly temperament theory constructs. The four temperaments have received various labels over time.[4] Because our focus in this paper is on applying temperament theory in research, especially via the KTS, we will adopt Keirsey's terminology. Keirsey labeled the four temperaments as: (1) Artisan, (2) Guardian, (3) Idealist, and (4) Rational. In the next section, an overview of the major defining characteristics of the temperaments is presented. These definitions are intended to present the accounting researcher with an overview of the kinds of characteristics that the KTS can provide for experimental purposes; they are not comprehensive. For detailed presentations see Keirsey (1998) and the Keirsey website (2002).

The Four Temperaments

Individuals with ***dominant Rational temperaments*** tend to be abstract and theory-oriented in their cognitive aspects, yet pragmatic and utilitarian in applying their ideas. They can be highly accomplished in strategy and in depth analysis. In terms of organizations, they are strong at marshalling resources and planning; in terms of applications, they are skilled at inventing and configuring. They take pride in being competent in their actions. They respect themselves for their ability to be autonomous and feel confidence because of their strong willpower. They tend to view the search for knowledge as a defining aspect of their personality. Thus, they trust in reason and are high achievers. Although an unusually small part of the population (5% to 7% according to the Keirsey website (2002)), they are frequently found in the sciences, technology and systems work.

Individuals with ***dominant Idealist temperaments*** tend to be cognitively abstract and theoretical, like those with Rational temperaments. However, the Idealist approach to applying and implementing their ideas is more group-oriented and cooperative than that of Rationals. Idealists tend to be skillful at diplomacy and integrating the contributions of others into their own work. Thus, they make excellent mentors and are often found among teachers and counselors. They also are outstanding leaders and high-level managers. They value highly social development and view interpersonal integration as a primary defining characteristic of their personality. Accordingly, they take pride in being ethical, empathetic and respectful of others. They tend to view the search for a unique identity, along with meaningful relationships, of primary importance. They are only slightly more frequent than Rationals, making up between 8% and 10% of the population (Keirsey website, 2002).

Individuals with ***dominant Artisan temperaments*** tend to be concrete, down-to-earth thinkers, with practical and utilitarian approaches to implementing goals. They can be extremely skillful in day-to-day operations and tactics. Thus, they are proficient at promoting and expediting plans. They also are skillful at composing and improvising. To Artisans, efficiency of action and elegance of solution tend to be as important as effectiveness and accuracy. They frequently measure their self-esteem in terms of gracefulness in action, degree of daring, and adaptability to change. According to the Keirsey website (2002), they comprise between 35% and 40% of the population. They gravitate towards careers in the arts and crafts, vocations requiring techniques, and operations work.

Individuals with ***dominant Guardian temperaments*** are also concrete, down-to-earth thinkers, similar to those with Artisan temperaments. However, Guardians tend to be more group-oriented in implementing and applying their ideas than are Artisans. They are thus highly proficient in logistics, with particular skills in supplying, conserving and protecting. They also display expertise in administration, supervising, auditing and inspecting. They pride themselves in being reliable and dependable in action, and valuing respectability highly. Security is central to their self-image, whether providing security, in the sense of being responsible for the goods of others, or receiving security, in the sense of being trusted and accepted by others. They are frequently found in careers involving material work and logistics, commerce and regulations. According to the Keirsey website (2002), they are the most numerous of the four temperaments, comprising from 40% to 45% of the population.

Keirsey, who developed the KTS, further defines the four temperaments to include concepts from linguistics and anthropology (Keirsey, 1991, 1998). From linguistics, Keirsey distinguishes those who prefer abstract communications

from those who prefer concrete communications. This aspect of the temperaments is measured by observing the types of words used. From anthropology's emphasis on the human reliance on tools, Keirsey distinguishes cooperative tool users from utilitarian or pragmatic tool users. This aspect is measured by observing the use of tools ranging from a simple one, such as a hammer, to the complex, such as a computer.

The intersection of the communications and tools aspects gives rise to a 2×2 four-cell matrix. Keirsey (1991, 1998) identified the resulting four combinations with the four temperaments as follows:

- Concrete communicator, pragmatic tool-user = Artisan
- Concrete communicator, cooperative tool-user = Guardian
- Abstract communicator, pragmatic tool-user = Rational
- Abstract communicator, cooperative tool-user = Idealists

Language and tool usage are generally accepted as being activities that distinguish human behavior from other types of behavior. These are activities that all people engage in on a routine basis, yet complex level. These aspects of human behavior can be reasonably used to distinguish between personalities. Thus, by using these two fundamental aspects of human behavior, Keirsey provides a scientific framework for understanding and justifying temperament theory's division of personality into four kinds. Previously, this four-way division of personalities was primarily based on psychological observation and philosophical speculation. By adding results from the sciences of anthropology and linguistics, Keirsey places temperament theory and the KTS on a more solid foundation, especially important for scientific researchers.[5]

Keirsey provides the experimental user of the KTS another invaluable service by incorporating personality constructs from the Jungian/MBTI personality type theory into the KTS. To examine this complex relationship between the temperament/KTS theory and Jungian/MBTI theory, Jungian personality type theory is discussed in the next section. This is followed by a section on the relationship between these two personality theories. The subsequent discussion then provides an in depth examination of the properties of the KTS as an instrument for measuring personality characteristics.

JUNGIAN PERSONALITY TYPE THEORY

The relationship between temperament theory and Jungian personality type theory is complex. There is general acceptance that the two theories are distinct yet with considerable overlap (Keirsey, 1998; Myers et al., 1998). Temperament theory is not, however, a derivative of Jungian theory, a mistake easily made when

one considers the similarities between the KTS and MBTI. Furthermore, both theories fall within the set of psychology theories known as personality theories (Feist & Feist, 1998; Hergenhahn & Olson, 1999). Beyond these commonalities, however, the two theories diverge. Temperament theory uses a model based on a four-way division of kinds of personalities (i.e. the four temperaments); Jungian theory uses a model of 16 types.[6] It is necessary, nevertheless, to discuss Jungian theory in regard to the KTS because the personality-trait variables used in the KTS (viz., Extraversion, Introversion, Sensing, Intuition, Thinking, Feeling, Judging and Perceiving) are directly derived from Jungian theory.

C. J. Jung's theory of personality types was developed in his 1921 work, *Psychological Types,* in response to his break with Freud over the latter's emphasis on the individual's unconscious (Jung, 1971; Hall & Nordby, 1973). Jung's approach has a strong information processing element, focusing on decision-making and the effect of personality on understanding of the world. Because of this orientation, Jungian theory shares many concerns and features with the currently dominant cognitive science model of psychology (Flanagan, 1991). Both approaches see information processing as a primary process or function of the mind. Cognitive psychology is based on a computational model of information processing developed from mathematical theory (Sipser, 1997). This approach, which also underlies computer science, tends to de-emphasize the role of personality in information processing. Research into the heuristics and biases used by humans, however, suggests that a predominantly computational model does not adequately explain the processes involved in human information processing (Kahneman et al., 1982; Baron, 1994). Jung's theory, with its inclusion of both information processing and personality traits, provides theoretical cohesion for the scattered results of heuristics and biases research and can supplement the cognitive science model of the mind with a needed synthesis of information processing and personality.

Jung's (1971) original personality type theory posits six personality traits which an individual uses to process information from the world. These six traits are grouped into three bipolar scales (i.e. complementary, dichotomous pairs). Four of the six traits derive from two fundamental mental processes (*Judging* and *Perceiving*) while the remaining two traits reflect an overall attitude toward the world (*Extraversion* and *Introversion).* Accordingly, the two fundamental mental processes (Judging and Perceiving) result in four mental functions – *Sensing, Intuition, Thinking* and *Feeling.* Individuals constantly use these four basic mental functions to perceive the world and construct their world-views. Sensing and Intuition are perceiving functions and deal with the type of inputs used for mental processing. Thinking and Feeling are Judging functions and transform the Sensing or Intuition inputs into each individual's unique world-view.

Furthermore, all individuals have relative interests in, or attitudes toward, the inner and outer worlds that affect their mental processes – *Extraversion* and *Introversion.*

Each individual is a mix of the six personality traits. However, Jung's theory postulates that, from each of the three pairs (Sensing or Intuition, Thinking or Feeling, and Extraversion or Introversion), one trait will be dominant or preferred. The resulting set of three *preferred* traits determines the personality type of the individual. Furthermore, one of the four mental functions will be dominant (Sensing, Intuition, Thinking or Feeling) leading to a dominant mental process (Perceiving or Judging). That is, a person will be dominantly a Judging-preference individual (Thinking or Feeling) or a Perceiving-preference individual (Sensing or Intuition). The three non-preferred traits are still present but have secondary roles in the nature of the personality. (By analogy, people have a preference for using either the right hand or left hand, and are therefore described as right-handed or left-handed. Nevertheless, everyone has some capability for using the non-preferred hand.) The resulting combination of preferred traits yields eight personality types as shown in Table 1.

Table 1. Jung's Eight Personality Types.

Three Trait Designation	Dominant Mental Function	Eight Personality Types
EST ENT	Judging (T or F)	Extraverted, dominant Thinking: The Extraverted Thinking type
EST ESF	Perceiving (S or N)	Extraverted, dominant Sensing: The Extraverted Sensing type
ESF ENF	Judging (T or F)	Extraverted, dominant Feeling: The Extraverted Feeling type
ENT ENF	Perceiving (S or N)	Extraverted, dominant Intuitive: The Extraverted Intuitive type
IST INT	Judging (T or F)	Introverted, dominant Thinking: The Introverted Thinking type
IST ISF	Perceiving (S or N)	Introverted, dominant Sensing: The Introverted Sensing type
ISF INF	Judging (T or F)	Introverted, dominant Feeling: The Introverted Feeling type
INT INF	Perceiving (S or N)	Introverted, dominant Intuitive: The Introverted Intuitive type

Notes: The following standard abbreviations apply:
E: Extraversion; F: Feeling; I: Introversion; N: Intuition; S: Sensing; T: Thinking.

When interpreting these personality traits, Jungian theory stipulates that no value judgments be made. This stipulation has two notable implications. First, neither aspect of a bipolar pair of traits is better or worse than the other. For example, being Extraverted is not better than being Introverted. Second, the fact that a trait is preferred (or non-preferred) does not imply how good someone is at using that trait. One may prefer to use the Extraverted trait yet be inept at Extraverted behavior. Conversely, one may have Introversion as a non-preferred trait, yet be very comfortable at acting Introverted. Jungian theory also states that the preferred traits are fixed in the individual from birth or at a very early age, with little subsequent change occurring (Pascal 1992).[7]

Mental Functions

A strong information processing orientation in Jung's theory is apparent in the mental functions. In information systems terms, the mental functions are inputting and processing in nature. However, the theory contains an equally strong behavioral aspect. Perceiving and Judging influence observable behavior. Perceiving, for example, determines which information a person actively gathers as relevant (preferred) in a situation. This information includes both internal and external types of data; i.e. things, people, events, feelings and ideas. Judging takes the information input from the preceding perceptions, and subsequently processes, organizes and transforms it into conclusions. These conclusions range from solutions to particular problems to the development of an integrated, holistic world-view. Note however, that people with different personality types may arrive at similar conclusions and behaviors. This is especially true for scenarios with specific, objectively definable and well-structured problems. But even in these scenarios, the data collection and processing may be markedly different among varying personality types.

Sensing and Intuition are theorized to be two styles of Perceiving, which may be respectively characterized as a preference for seeing either "the trees" or "the forest." These are input functions that describe sources of data to be subsequently processed by the judgment functions. Sensing-preference individuals prefer direct or objective perceptions made through the basic senses; i.e. hearing, sight, etc. Individuals oriented toward Sensing are inclined to center on immediate experience because the senses gather data from events presently happening. They tend to be practical, matter-of-fact, realistic, observant, and pragmatic, and have a good memory of facts.

Intuition-preference individuals prefer subjective perceptions of possibilities, structures, and meanings over sense-based perceptions. They observe associations among objects, experiences, ideas and entities. Individuals who prefer this

function tend to see the overall picture and are imaginative, speculative, abstract and inventive. Such perceptions are often made through insight. Intuitive perception includes concepts and relationships beyond the capabilities of the senses. Intuition should not be equated with vague feelings, since it can be rigorously logical.

Thinking and Feeling are the two types of functions from the Judging bipolar scale. These are processing functions that transform the inputs from the perception functions. Thinking-preference individuals prefer connecting thoughts and experiences together logically. They tend to be analytical and objective, and focus on causal associations. They are especially concerned with principles such as fairness and justice, which may make them seem impersonal when dealing with others. Individuals who prefer Thinking are also frequently systematic problem solvers.

Feeling-preference individuals tend to rely on values when involved in decision-making, either personal or group. Thus, they tend to develop characteristics such as a concern for the human side of problems, sympathy and compassion. A general understanding of people may make them seem to be overly tender hearted. Those with this approach to making judgments are usually more attuned than Thinking-preference individuals to the desires, values and needs of others.

While the four functions may be divided according to whether they are perception-input or judgment-processing, they also compete with each other to provide the mind with a preferred objective or viewpoint in understanding the world. Each function provides a different objective to conscious mental activity. Sensing offers to the mind a full experience of the immediate and real sensory world. Intuition offers a deep comprehension of the structure and possibilities of the world, both imaginative and sensory. Thinking offers the mind a logically and rationally ordered world. Feeling offers a rational world also, but one built around an orderly arrangement of subjective values. These conflicting goals interact with each other so that only two will be primary in the individual, one from each of the two mental bipolar scales. The relative importance of these two preferred mental functions is determined by their relationship with the individual's attitude toward the world.

Attitudes Towards Internal and External Aspects of the World

According to Jung's theory, Extraversion and Introversion are fundamental attitudes that describe the individual's approach toward the world in its internal and external aspects. Jung believed that this distinction was the most fundamental in the personality of the individual (Jung, 1971). While a healthy life requires a mix of both attitudes, one clearly dominates in the personality.

Extraversion is an attitude wherein the individual's focus is on people and objects of the external world. Extraverts are keenly sensitive toward the environment. For Extraverts, external objects – personal and impersonal – are sources of energy and direction. They focus their particular Perceiving-functions (Sensing-Intuition) and Judging-functions (Thinking-Feeling) on things in the outside environment. Extraverts receive the energy for these mental processes from experiences involving external activities and events. Extraverts tend to be action-oriented, expressive, and outgoing. They excel at oral communication and have learning-styles that are group-based and action-based.

Introversion is an attitude wherein the individual's attention is directed primarily to the inner environment of the mind and those things that comprise this subjective world of thoughts and feelings. They get their energy from the activities of the mind; i.e. thinking, feeling, and reflecting; and they direct their energy and actions toward this realm. They focus their particular Perceiving-functions (Sensing-Intuition) and Judging-functions (Thinking-Feeling) on things in the internal environment. They put their energy into concepts and ideas, committed to making them clear and accurate. Those who prefer Introversion tend to be contemplative and detached from the external world. They enjoy and need solitude and privacy. This is when they re-energize. They prefer written communication and have individual learning-styles.

Developing a Fourth Bipolar Scale: Orientation towards the External World

Jung theorized that between the two preferred functions from the two mental bipolar scales, one function would be dominant. This dominance defines whether a person is primarily interested in data collection or data processing. When the Perceiving bipolar scale (Intuition-Sensing) is dominant, one is more interested in gathering information; when the Judging bipolar scale (Thinking-Feeling) dominates, one is more concerned about solving problems and arriving at conclusions. A review of Table 1 reveals that the non-dominant yet preferred mental function was essentially "dropped" from Jung's description of the personality (Myers & Myers 1995). For example, an EST who is dominantly a judging-preference individual is described as "an extraverted thinking type"; i.e. the preferred Sensing function is not included in the description. Jung believed that this "dropped" mental function, referred to as the auxiliary function, was the least differentiating of the functions in the personality, primarily because it functioned unconsciously.[8]

I. B. Myers modified Jung's theory further by developing a fourth personality bipolar scale. (See Myers & Myers, 1995 for a discussion of the history of this development.) This new scale was introduced in response to mixed

results from empirical work attempting to test Jung's theory (Meier & Wozny, 1978; Rosenak & Shontz, 1988). These results indicated that Jung's eight personality types were not sufficient to classify the data concerning personality characteristics. The new bipolar scale was introduced to clarify ambiguities about the interrelationship or interaction effect among the preferred functions in the mental bipolar scales. Specifically, it was developed to indicate whether an individual shows an overall preference for the preferred function in the Judging bipolar scale (Thinking-Feeling) or the preferred function in the Perceiving bipolar scale (Sensing-Intuition) *when dealing with the external world.* One result of introducing this fourth bipolar scale was to make the auxiliary function an important aspect of the personality type.

This new bipolar scale from I. B Myers, often referred to as an attitude or orientation to the external world, consists of two traits, *Judging* and *Perceiving.* The Judging and Perceiving traits in the fourth bipolar scale may initially appear to be present in Jung's theory. To some extent this is true. As discussed above, the Judging mental function contains Feeling and Thinking functions, and the Perceiving mental function contains Sensing and Intuition functions. However, the new (fourth) bipolar scale is an attitude that indicates one's preference for either the Judging or Perceiving mental function in relation to the external world, regardless of whether one is otherwise an Extravert or Introvert. That is, what the fourth bipolar scale is trying to ascertain is whether the individual prefers using one of the Judging mental functions (Thinking or Feeling) or one of the Perceiving mental functions (Sensing or Intuition) when dealing with the outside world. The terminology overlaps and may seem confusing (see note 2).

Judging-preference individuals are those who prefer to use the Judging mental function when dealing with the outer world. Judging-preference individuals prefer planning and organizing activities. They like to see problems solved to completion. They especially desire closure in their mental activities and are, therefore, thorough, conscientious, methodical and decisive. Perceiving-preference individuals are those who prefer to use the Perceiving bipolar scale when dealing with the external world. They prefer openness in their mental activities and are concerned that all possible information is gathered. They like to keep the problem open as long as possible. This tends to make them spontaneous, curious and adaptive decision makers. From an information processing perspective, Judging-preference individuals may be characterized as individuals who like to perceive the output as completed and closed to further revision. Perceiving-preference individuals prefer to see the output as incomplete and open to additional processing.

In summary, Jungian personality type theory, postulates eight personality traits from four bipolar scales: Extraversion-Introversion; Sensing-Intuition;

Thinking-Feeling; Judging-Perceiving. All eight traits are present in each person and can be treated as continuous variables. This continuous approach is not, however, the usual interpretation attached to the presence of the traits in individuals. The traits are generally treated as dichotomous and exclusive. In an individual, one trait from each of the four bipolar scales is preferred. According to Jungian theory, although individuals have all eight traits to some extent, the four preferred traits primarily define the personality. This system of preferences results in 16 possible combinations of the four preferred traits, which are referred to as the 16 Jungian personality types. As will be discussed in the next section, this system of four bipolar scales is used in the KTS (Keirsey, 1998). Table 2 provides general characteristics and occupational tendencies of the 16 Jungian personality types.

The Relationship between Temperament Theory and Jungian Theory

As noted earlier, temperament theory is not an offshoot of Jungian personality type theory. Nor does Jungian theory derive from temperament theory, although the latter predates the former and influenced Jung. Yet, the above excursion into Jungian theory is necessary in order to understand Keirsey's formulation of temperament theory and the Keirsey Temperament Sorter because Keirsey uses the eight Jungian personality traits in his system. His justification for doing so is primarily observation-based, not theory-driven (Keirsey, 1998; also see note 5 on this aspect of personality theories in general). Although Keirsey believes that Jungian theory is "cumbersome and self-contradictory" (Keirsey, 1998, p. 15), he accepts the constructs underlying the MBTI as valid and reliable. According to Keirsey (1998), the individual traits or variables in the MBTI (i.e. Extraversion, Introversion, Sensing, Intuition, Thinking, Feeling, Judging and Perceiving) accurately measure characteristics of the personality, but he disagrees with how MBTI/Jungian theory subsequently combines and interprets these individual variables (discussed in detail later).

KTS temperament theory is based on a systems model, focusing on the configuration of the whole. MBTI Jungian theory, on the other hand, is based on a dynamic parts model, in which the attitudes and functions can be consciously manipulated. Keirsey and Myers disagree in what constitutes similar personalities, primarily because of their different categories for grouping the various personality traits. Myers groups traits according to Jungian type theory, while Keirsey groups traits according to temperament theory. Divergences in Keirsey's and Myers' "similar types" have been empirically tested and are discussed below (Hobby et al., 1987).

Table 2. The 16 Personality Types with General Characteristics and Occupational Tendencies.

	Sensing (S)		Intuition (N)		
	Thinking (T)	**Feeling (F)**	**Feeling (F)**	**Thinking (T)**	
Judging (J)	**ISTJ** Practical, sensible, decisive, logical, detached. *Management and administration.*	**ISFJ** Practical, concrete, cooperative, sensitive. *Education, health care, and religion.*	**INFJ** Insightful, symbolic, idealistic, committed, compassionate. *Religion, counseling, and teaching.*	**INTJ** Insightful, long-range thinkers, clear, rational, detached. *Science, computers, and technical fields.*	**Introversion (I)**
Perceiving (P)	**ISTP** Detached, logical problem solvers, pragmatic, factual. *Skilled trades and technical fields.*	**ISFP** Trusting, kind, sensitive, observant, practical, concrete. *Health care and business.*	**INFP** Sensitive, caring, idealistic, curious, creative, visionary. *Counseling, writing, and arts.*	**INTP** Logical, curious, detached, insightful, contemplative. *Scientific and technical fields.*	
Perceiving (P)	**ESTP** Observant, active, rational problem solvers, assertive. *Marketing, business, and skilled trades.*	**ESFP** Observant, specific, active, sympathetic, idealistic, warm. *Health care and teaching.*	**ENFP** Curious, creative, energetic, friendly, cooperative, warm. *Counseling, religion, and teaching.*	**ENTP** Creative, imaginative, theoretical, analytical, rational, questioning. *Science, management, and technology.*	**Extraversion (E)**
Judging (J)	**ESTJ** Logical, decisive, objectively critical, practical, systematic. *Management and administration.*	**ESFJ** Factual, personable, cooperative, practical, decisive. *Education, health care, and religion.*	**ENFJ** Compassionate, loyal, imaginative, likes variety, supportive. *Arts, religion, and teaching.*	**ENTJ** Analytical, assertive, conceptual thinkers, innovation planners. *Management and leadership.*	

Notes: General personal characteristics are shown in regular font style; occupational tendencies are shown in italics.
Source: Reproduced from Wheeler (2001: 128) with permission.

Keirsey (1998) uses the Jungian-based variables to arrive at operationalized definitions of the four temperaments; i.e. definitions that may be equated with measurements from the KTS. Specifically, he equates the four temperaments with 2-way interactions or combinations of the MBTI/Jungian variables. Thus, an SP (Sensing-Perceiving) combination indicates an Artisan temperament, an SJ (Sensing-Judging) combination indicates a Guardian temperament, an NF (Intuition-Feeling) combination indicates an Idealist temperament, and an NT (Intuition-Thinking) combination indicates a Rational temperament. The KTS, not the MBTI, is used to measure these traits in the subjects for arriving at the temperaments because the KTS operationalizes these traits slightly differently than does the MBTI (i.e. the questions in the two instruments are different). Keirsey is only interested in using the theoretical constructs (i.e. the eight traits) underlying the MBTI, not the MBTI instrument per se. Keirsey contends that these combinations of variables derived from the KTS identify individuals with the characteristics of the corresponding temperament.[9] Table 3 presents the four temperaments in a grid using the eight Jungian/MBTI personality traits.

Some discussion of the relationship between the 16 MBTI types and the four KTS temperaments is required in order to understand how the KTS can be used for research. Temperament theorists, including Keirsey (Keirsey, 1998; Keirsey website, 2002), typically group the 16 Jungian/MBTI types into four groups corresponding to the temperaments. This grouping is done to show the degree of overlap between the two personality theories and to facilitate movement between the various classification schemes. Table 4 provides an example of this approach based on Keirsey's approach (1991, 1998).

As depicted in Table 4, the current names of the Keirsey temperaments are presented with correlation to the MBTI traits. As noted above, both KTS instruments and the MBTI use four bipolar scales and eight traits. There are, therefore, 16 *types* in both Keirsey's and Myers's schemes. However, the two sets of 16 types are not the equivalent. Differences between the two schemes arise, not so much from the measurements made by the respective instrument, but from the descriptions attached to the measured variables. Users mistakenly assume they can interchange Keirsey's type descriptions for the same type letters of the MBTI; while this is frequently done, it is not advised, due to validity concerns pertaining to differing theoretical origins. Keirsey has emphasized that making his 16 types "conform to ancient temperament theory took juggling" (Frisbie, 1988).

In summary, Keirsey uses the eight Jungian/MBTI traits as variables in the KTS and temperament theory because they are effective at measuring detailed characteristics of the personality. He then assembles and interprets the data from these measurements in a manner different than that done by Jungian and MBTI

Table 3. The Four Temperaments with General Characteristics.

	Sensing (S)		Intuition (N)		
	Thinking (T)	Feeling (F)	Feeling (F)	Thinking (T)	
Judging (J)	**SJ** **Guardian temperament** Concrete, cooperative, logistical, dutiful, pessimistic, authoritative, and beneficent. *Administrator, auditor, and judge.*		**NF** **Idealist temperament** Abstract, cooperative, diplomatic, altruistic, credulous, benevolent, and intuitive. *Teacher, ambassador, and doctor.*	**NT** **Rational temperament** Abstract, utilitarian, strategic, pragmatic, skeptical, autonomous, and reasonable. *Inventor, architect, and scientist.*	**Introversion (I)**
Perceiving (P)	**SP** **Artisan temperament** Concrete, utilitarian, tactical, practical, optimistic, audacious, and impulsive. *Salesperson, airline pilot, and entertainer.*				
Perceiving (P)					**Extraversion (E)**
Judging (J)	**SJ** **Guardian temperament** (continued)				

Notes: The Guardian temperament is shown in two sections because of the table format being based on the four bipolar scales. It is, nonetheless, a single temperament like the other three.
General personal characteristics are shown in regular font style; occupational or role tendencies are shown in italics.
Sources: Myers et al. (1998) and Keirsey (1998), with modifications.

Table 4. The Four Keirsey Temperaments and 16 Types, Referencing Correlated MBTI Traits.

	Idealists NF	**Rationals NT**	**Guardians SJ**	**Artisans SP**
Directive role*	**Mentors NFj**	**Organizers NTj**	**Monitors SJt**	**Operators SPt**
Extroverted **(e)**	Teacher eNFj	Field Marshal eNTj	Supervisor eSJt	Promoter eSPt
Introverted **(i)**	Counselor iNFj	Mastermind iNTj	Inspector iSJt	Crafter iSPt
Informative role*	**Advocates NFp**	**Engineers NTp**	**Conservators SJf**	**Players SPf**
Extroverted **(e)**	Champion eNFp	Inventor eNTp	Provider eSJf	Performer eSPf
Introverted **(i)**	Healer iNFp	Architect iNTp	Protector iSJf	Composer iSPf

Notes: Keirsey's terms for the 16 types have evolved. The 16 types presented here are the most recent nomenclature. Myers-Briggs terms have been inserted to indicate correlation, not identity.
* This grouping stems from the kinds of relationships the temperaments are willing to have with others. Typically, roles are seen as primarily directive (i.e. influencing the actions of others) or informative (i.e. providing others with ideas or data) (Keirsey, 1991).
Sources: Keirsey (1998) and Keirsey (1991), with modifications.

theorists. A rough analogy might be one using the inch as a measurement but then aggregating it in groups of ten (as with the metric system) instead of groups of twelve (as with the U.S. customary system).

THE KEIRSEY TEMPERAMENT SORTER

The Keirsey website (2002) makes the claim that the KTS is "the No. 1 online personality test" and that this test is "used by many Fortune 500 companies to test their employees and by major Universities to test their students" (Keirsey website, 2002: Introduction screen). Furthermore, the KTS is used in accounting and business research that has resulted in publications and dissertations (e.g. Gul & Fong, 1993; Hozik & Wright, 1996; Swanger, 1998).

The Keirsey Temperament Sorter (KTS) is a forced-choice test. The forced-choice format consists of a question followed by two responses representing the two traits in one of the four bipolar scales. The initial version was developed and included in Keirsey's 1978 book. The more recent version is referred

to as KTS-II and consists of 70 questions. The KTS-II is available both in a paper format (Keirsey, 1998) and online (Keirsey website, 2002). Both versions of the KTS, similar to the multiple versions of the MBTI, measure the eight Jungian-based traits (i.e. Extraversion, Introversion, Sensing, Intuition, Thinking, Feeling, Judging and Perceiving) and indicate an individual's four preferences from these pairs. However, the resulting personality description from the KTS is based on temperament theory, not Jungian type theory. Temperament theory differs from Jungian type theory underlying the MBTI in that the former places emphasis on the 2-way interactions of certain preferences (Keirsey, 1998; Myers et al., 1998).

Two profiles may be derived from taking and scoring the KTS. The first profile is general to each of four temperaments (SJ, SP, NF, NT); detailed 16 type profiles are also available. Keirsey's book, *Please Understand Me II* includes both temperament and type profiles (Keirsey, 1998).[10] While the KTS uses the same four Jungian-based bipolar scales and accompanying eight traits as the MBTI, the personality profile descriptions provided by the two instruments differ. As will be discussed below, convergent validity studies indicate that KTS and MBTI profiles are, nevertheless, significantly correlated (McCarley & Carskadon, 1986; Hobby et al., 1987).

Validity and Reliability of the Keirsey Temperament Sorter

Issues concerning the validity and reliability of the Keirsey Temperament Sorter are critical because the validity and reliability of an instrument determine its usefulness experimentally. The authors located seven studies examining convergent validity issues surrounding the KTS. Of these, four investigate the relationship between the Keirsey theory and the MBTI (Hobby et al., 1987; McCarley & Carskadon, 1986; Ruhl & Rodgers, 1992; Ware & Yokomoto, 1985). The other three are empirical studies comparing the KTS and the MBTI (Kelly & Jugovic, 2001; Quinn et al., 1992; Tucker & Gillespie, 1993).

To investigate the Keirsey theory, Ware and Yokomoto (1985) first administered the MBTI to undergraduate psychology students to identify their personality traits. Each subject was then given a packet containing several different personality descriptions based on Keirsey's theory, including the subject's type, reversed function type, reversed attitude type, and the opposite type. Next, the subjects were asked to indicate the relative accuracies of the different personality descriptions. Ware and Yokomoto (1985) found that subjects perceived the Keirsey-descriptions of their MBTI-indicated personality type to be more accurate than the other type Keirsey-descriptions (i.e. those not corresponding to the MBTI indicated personality type). Since all of these

descriptions were based on Keirsey's theory, the researchers interpreted this result as supporting the convergent validity of Keirsey's theory, using the MBTI as the validity criterion.

McCarley and Carskadon (1986) used procedures similar to Ware and Yokomoto's (1985) except that they provided the subjects with individual elements from Keirsey's descriptions instead of the whole type descriptions. McCarley and Carskadon (1986) also provided subjects with elements from type descriptions based on the MBTI theory. They found that subjects perceived the elements of Keirsey's descriptions similar to the elements from the MBTI theory.

A replication of McCarley and Carskadon (1986) found similar ratings by participants in overall accuracy of personality descriptors between Keirsey and the MBTI (Ruhl & Rodgers, 1992) that support the earlier study. There was only one significant difference that emerged, Thinking vs. Feeling. Of the sixteen types studied, two of the three types that preferred Keirsey's descriptors were Thinking types, while all three that preferred MBTI were Feeling types.

After determining subject's personality types by administering the MBTI, Hobby et al. (1987) provided each subject with two complete descriptions of the subject's indicated personality type, corresponding to the "most similar" type per Keirsey and per Myers. Note that the type descriptions provided to participants were not of the types as indicated by the MBTI, but were instead what each subject would have perceived as "most similar" to the scored type. The results found for most types no significant difference in how subjects evaluated the relative accuracy of the two type descriptions. In a few types, the MBTI-based descriptions were perceived to be more accurate. However, the results of this study are somewhat ambiguous in that the authors did not use the type descriptions in their original form.

Overall, the studies by Hobby et al. (1987), McCarley and Carskadon (1986), Ruhl and Rodgers (1992) and Ware and Yokomoto (1985) are supportive of the convergent validity of the Keirsey Temperament Sorter. However, two qualifications are necessary to this conclusion. First, these studies investigate only the theory (Keirsey's temperament theory) underlying the instrument (KTS), not the instrument per se. In none of the studies was the KTS used. Second, the studies used the type descriptions from Keirsey's work, not the temperament descriptions; i.e. the two-way interactions were not examined.

The following three studies used the KTS instrument. Kelly and Jugovic (2001) took concurrent measures from undergraduate students from the KTS-II and the MBTI Form G. They found correlations between individual KTS and MBTI traits ranging from a low of 0.60 to a high of 0.78. According to Kelly

and Jugovic (2001, p. 55), these correlations are "moderate to strong" and "indicate that the KTS-II has satisfactory concurrent [i.e. convergent] validity." Using the earlier version of the KTS, Quinn et al. (1992) found correlations between KTS and MBTI traits ranging from 0.54 to 0.74 on a sample of business undergraduates. Tucker and Gillespie (1993), also using the previous version of the KTS, had correlations between the KTS and MBTI ranging from 0.68 to 0.84 on undergraduate psychology students. As with Kelly and Jugovic (2001), these last two studies provide moderate to strong support for the convergent validity of the KTS.

Another study investigated reliability issues of the KTS (Waskel 1995). Waskel (1995) empirically examined the internal consistency of the prior version of the KTS and found alpha coefficients of 0.74 for the Extraversion-Introversion scale, 0.89 for the Sensing-Intuition scale, and 0.87 for the Thinking-Feeling scale and 0.88 for the Judging-Perceiving scale. Alpha coefficients of 0.70 and higher are generally considered acceptable levels of instrument reliability (Nunnally & Bernstein, 1994).

In summary, numerous validity and reliability studies have been conducted on the KTS and its underlying temperament theory. The results consistently provide moderate to strong support for both the instrument and theory. Thus, researchers are justified in assuming that the KTS provides reliable measurements of certain psychological traits and that the descriptions (as derived from temperament theory) of these measured traits have valid content and discriminatory power. One weakness in the validity testing that should be noted is that little has been done using the specific two-way combinations that correspond to the four temperaments.

Comparing the KTS and the MBTI

Like the KTS, the MBTI is a forced-choice test (Myers et al., 1998). It is significantly longer than the KTS, consisting of 93 questions instead of 70, and thus requires more time to administer. The MBTI is copyrighted and can be purchased from Consulting Psychologists Press, Inc. (CPP) in self-scoring, computer mail-in scoring and online versions. The MBTI has evolved since 1942 to its present version (Form M). Form M replaced the prior Form G as of 1998. Both versions are still acceptable for use. CPP requires those administering the MBTI to meet certain psychological testing training and education requirements, which are not required to use the KTS.

The MBTI has undergone extensive reliability and validity testing (Harvey, 1996; Myers et al., 1998; Wheeler, 2001). Tests of internal consistency and temporal stability have consistently provided strong support for the reliability

of the scores from the instrument. Testing of discriminant, convergent and construct validity has not been as unvaryingly supportive of the MBTI as those from reliability studies. The results indicate that the MBTI is measuring aspects of personality in a way usually consistent with Jungian theory but is not capturing the personality in its entirety.

A common criticism of the MBTI and Jungian personality theory is that the former does not capture and the latter does not portray the personality in its entirety. This is a criticism valid for all personality theories, including temperament theory, and probably for any scientific model. Theories and models deliberately oversimplify. The crucial issue is not so much whether the theory captures all of major aspects of the phenomenon under investigation but whether it allows researchers to make discriminatory predictions that are testable. Undoubtedly, a cost-benefit tradeoff is involved; a balance must be struck between increasing understanding and furthering research. In this regard, the KTS and temperament theory are significantly different than the MBTI and Jungian theory, offering the researcher opportunities to make unique predictions. Temperament theory includes the temperaments as constructs not found in Jungian theory. As noted above, little validity testing has been done in this area of the KTS; therefore, this represents an area open to future research.

RESEARCH OPPORTUNITIES USING THE KEIRSEY TEMPERAMENT SORTER

Use of the Keirsey Temperament Sorter for research may be approached from two angles. First, the KTS may be used in a manner similar to the MBTI. The KTS, like the MBTI, results in preference scores for the four bipolar scales. Thus, for example, if one wants to use scores for the two mental bipolar scales to capture cognitive style (as done by Chenhall & Morris 1991 and Vassen et al., 1993), then these scores can be provided by either the KTS or the MBTI. The decision should be based on such considerations as reliability, validity, convenience, availability and cost. Second, as discussed above, unique aspects of temperament theory distinguish it from Jungian personality theory. If the research being undertaken involves temperaments, then the KTS, not the MBTI, should be used.

The KTS can be used to address three general accounting research areas. First, research can be done on college-aged accounting students; e.g. the distribution of indicated personality types and temperaments, and relationships of personality traits to academic performance and choice of major. Second, research can be conducted on teaching styles; e.g. understanding the interaction of teaching styles and learning styles from the respective personality traits

of teachers and students. Third, research in the accounting profession can be done; e.g. the distribution of indicated personality types and temperaments, and relationships of personality traits to career success and job-placement. The use of the KTS within businesses should also be examined since it is widely used for management purposes (Keirsey website, 2002).

Wheeler (2001) gives an extensive review of prior research using the MBTI to investigate the role of personality traits and types in accounting education and the profession. This review is suggestive of numerous specific opportunities for applying personality-based instruments to these three research areas. For example, despite changes in the accounting environment and repeated calls for more diversity in the profession, the distribution of indicated personality traits in the accounting profession has remained remarkably stable.

It appears that much of the stability in personality traits found in the profession results from the college process and environment. One study (Laribee, 1994) indicates that accounting education involves a filtering-out process that decreases the percentages of Extraversion, Intuition, Feeling and Perceiving-preferences among accounting students. Research to explain this filtering is needed. Is it related to teaching-style, although accounting teachers tend to be Intuition-preferenced? (Wolk & Nikolai, 1997). If this filtering is common, discovering ways to decrease it may allow for increased diversity among accounting students. At a time of concern about the future of accounting, this is an area of valuable research.

The learning styles of individuals affect their performance. Personality theory predicts that personality traits affect learning styles (Myers et al., 1998). Accounting researchers should therefore be able to use personality-based instruments to investigate this relationship among accounting students. Also, many other learner characteristics can be investigated using personality-based instruments; e.g. written vs. oral, team vs. individual, concrete vs. theoretical, and structured vs. open-ended problems (Myers et al., 1998).

Studies of the academic performance in undergraduate accounting courses have produced mixed results (Gul & Fong, 1993; Nourayi & Cherry, 1993; Oswick & Barber, 1998). Gul and Fong (1993) and Nourayi and Cherry (1993) found a correlation between personality traits and performance; Oswick and Barber (1998) did not. Further research may be conducted to examine why these mixed results occurred; e.g. due to differences in the samples. Also, future research should investigate why the relationship between personality traits and performance is weaker in introductory accounting courses than in later accounting courses. If it is not performance, then what personality traits determine the self-selection of those proceeding to the later accounting courses and becoming accounting majors and accountants?

Research is also needed on the personality traits, types and temperaments of accounting faculty. Investigation of whether there is a relationship between accounting faculty personality and their effectiveness with certain teaching methods is warranted. Other potential areas include the relationship between accounting faculty personality traits and learning styles of students, and the relationship between accounting faculty personality traits and the accounting courses they teach.

Only two studies have been done in the area of how accountants perform their professional tasks (Chenhall & Morris, 1991; Vassen et al., 1993). These studies looked only at auditors, cognitive processes, and the Sensing-Intuition and Thinking-Feeling scales as non-interactive variables. Research of other professional areas in accounting (tax, consulting, and managerial) is of interest. A strength of personality-based research compared to alternative types of psychology-based research is that aspects of the mind besides cognition and information-processing can be examined; i.e. the mental functions as separate variables. Thus, research using two-way, three-way and four-way preference interactions needs to be done in the various aspects of the profession.

Team or group dynamics is one of the more promising areas of personality-based research (Hammer & Huszczo, 1996). According to the AICPA's Vision (1999), the ability to work in teams, in particular "within diverse, cross-functional teams" is a necessary accounting skill. Team formation is one of the major areas where personality-based instruments are used in the workplace. Research as to the effectiveness of these instruments in this area is needed. Recall that personality-based instruments can examine both cognitive and non-cognitive aspects of teams. Cognitively, teams may be investigated as problem-solving and decision-making units by looking at the distribution of mental functions among team members. Do certain mixes lead to greater efficiency or effectiveness in solving problems? Personality-based instruments are also capable of investigating numerous non-cognitive dimensions of teams. For example, because communication in the team affects overall team performance, are team members with certain combinations of traits better at communicating than others?

Research on team formation, as opposed to performance, using personality-based instruments is needed to investigate the effort management should put into deliberately constructing teams as opposed to allowing teams to self-select. If deliberately forming teams is necessary, then research should focus on which traits and trait interactions are important. Are some membership arrangements better for certain types of tasks than others? For example, research may examine the traits of effective teams as they vary with tax, auditing, financial, or consulting engagements. Similarly, research may indicate that the traits of the client also have an impact on the accounting team's effectiveness.

Furthermore, parallel research questions on teams in the classroom need to be answered using personality-based instruments. Can the accounting instructor improve the effectiveness of team learning and team projects by selecting membership based on personality traits, temperaments or types? Do teams self-select in an equally advantageous manner? Do the personality traits of effective teams varying with the type of accounting course? For example, are the characteristics and skills of effective financial accounting teams different than those of auditing teams?[11]

In addition to being used for capturing the eight traits and 16 personality types, the KTS can be used to specifically investigate hypotheses involving the four temperaments. As discussed above, temperament theory claims to capture important differences between personalities with the four temperaments based on two-way interactions of personality traits (Keirsey, 1998; Berens, 1996, 1998). Since temperament profiles are different from Jungian type profiles, different predictions concerning behavior are expected. Another advantage of the temperament theory approach is that it often requires a smaller sample size than the 16-personality type approach, depending on the hypothesis under consideration.

Numerous MBTI studies have been conducted that include two-way interactions of personality traits among the various combinations of traits examined. These include research into personality characteristics, psychotherapy, health, education, careers, leadership, management and teams (DiTiberio, 1996; Haley, 1997; Hammer, 1996; Kirby & Barger, 1996; Myers & McCaulley, 1985; Myers et al., 1998; Quenk & Quenk, 1996; Shelton, 1996; Walck, 1996). These are not temperament studies per se since they do not look exclusively at the four two-way combinations corresponding to the temperaments. Nor do they employ temperament descriptions in interpreting the results. They do, nevertheless, indicate that there is a great deal of potential for temperament studies because the temperament two-way combinations are among the interactions studied.

CONCLUSIONS

In this paper, three points have been emphasized about temperament theory and the KTS, especially in comparison to Jungian theory and the MBTI. First, temperament theory is unique. Second, the KTS has tremendous potential as a research instrument, similar to the MBTI. Third, while more validity and reliability testing of the KTS is warranted, much has already been done.

In relation to the uniqueness of temperament theory, temperament theory is different from Jungian/MBTI theory. Temperament theory has a different, pre-Jungian history and views Jungian theory as coming, somewhat indirectly, from

temperament theory. There is significant overlap between the two theories. Both theories and their related instruments use the eight Jungian/MBTI personality traits (Extraversion, Introversion, Sensing, Intuition, Thinking, Feeling, Judging and Perceiving) and the concepts of preference and dominance.

The MBTI and KTS that accompany the two theories are similarly related. They both measure the four bipolar scales and accompanying eight traits. However, the KTS was developed specifically to capture the four two-way interactions of the bipolar scales found in temperament theory (i.e. the four temperaments). This is not true of the MBTI, which was developed to capture the four-way interactions (i.e. the 16 types).

The fact that the two instruments employ the same variables to measure the same traits allows researchers to use results from both streams. Although both instruments use the same four scales, how the two theories define the scales are slightly different. Both theories operationalize the same constructs but with slightly different emphases. For example, the KTS "Sensing-Intuition" and the MBTI "Sensing-Intuition" are operationalizations of the same theoretical construct (Sensing-Intuition). The questions used in the two instruments to get at this trait are different, as are the ultimate applications; i.e. determine four temperaments vs. 16 types. This difference is understandable given the highly inclusive nature of the underlying constructs. Personality theorists are trying to capture a very complex phenomenon (the personality) with only four bipolar scales. This partially explains the less than perfect correlation between the two instruments, generally around 0.75.

In relation to KTS' potential as a research instrument, the KTS is a psychometric instrument used to measure constructs (i.e. Extraversion-Introversion; Sensing-Intuition; Thinking-Feeling; Judging-Perceiving) deriving from personality type theory, like the MBTI. Accordingly, researchers may use these measurements in several different ways:

(1) To determine the four two-way interactions known as the temperaments. This is what the KTS was specifically developed for, yet little has been done in this area of accounting.
(2) To determine the four bipolar scales and accompanying eight traits, and use these separately. For example, the Sensing-Intuition and Thinking-Feeling scales have been used separately to operationalize "cognitive style" (e.g. in accounting, Chenhall and Morris 1991). Another example of a study using traits in the accounting profession is the use of the Extraversion-Introversion to predict CPA supervisory opportunities (Satava 1997).
(3) To determine the 16 MBTI/Jungian types. This is not the reason for which the KTS was developed, but the reason for the MBTI. Nevertheless, the

KTS can provide this. Because of the overlap between the KTS and MBTI, researchers using the KTS may draw from and expand on the results from the extensive MBTI research stream.

(4) To obtain multiple measures of the same phenomenon using both the KTS and the MBTI. Multiple measures yield different applications, which advances research and integration across models. Also, they assist individuals in determining their "best fit" type. Evidence of this is apparent in the literature as temperament theory is currently being fused with concepts of systems thinking and multiple intelligences.[12]

In summary, the Keirsey Temperament Sorter is an instrument that allows the accounting researcher to measure certain traits of the personality defined in terms of temperament theory and, indirectly, Jungian personality type psychology. These theories do not limit themselves to any one aspect of the personality, such as information processing, but instead look at the whole personality. The broad scope of temperament theory is an advantage because, through the use of the KTS, it enables the accounting researcher to capture more variation in experimental subjects than that of more specialized psychological paradigms and instruments. There is a rich stream of potential accounting research that can be conducted with the KTS, as illustrated by the small number of accounting articles using personality theories and instruments. The personality types, temperaments and traits of accounting students, teachers, and professionals have been investigated in only a few areas. The MBTI research stream is another rich source from which the accounting researcher can draw examples because the KTS and MBTI have many features in common. Accounting researchers have many opportunities open to them using the Keirsey Temperament Sorter and other Jungian-based instruments.

> It is not the purpose of psychological typology to classify human beings into categories – this in itself would be pretty pointless. Its purpose is rather to provide a critical psychology which will make a methodical investigation and presentation of the empirical material possible. First and foremost, it is a critical tool for the research worker, who needs definite points of view and guidelines if he is to reduce the chaotic profusion of individual experiences to any kind of order. (Jung, 1971, pp. 554–555).

NOTES

1. The literature uses the following standard abbreviations: E for Extraversion, I for Introversion, N for Intuition, S for Sensing, T for Thinking, F for Feeling, J for Judging, and P for Perceiving. For ease of reading, we have avoided using the abbreviations except where necessary. Accordingly, the abbreviations are used in tables and with combinations of traits (e.g. ESTJ for Extraversion-Sensing-Thinking-Judging). We have

retained the convention of capitalizing traits and functions to distinguish the specialized terms from their everyday counterparts – thus, Introversion (the trait) vs. introversion (the common term).

2. The terminology may seem confusing initially. *Traits* refer to individual characteristics such as Extraversion, Introversion, Intuition, Sensing, Thinking, Feeling, Judging or Perceiving. The *bipolar scales* (or *dichotomies*) are the four complementary pairs of the traits; e.g. the Intuition-Sensing pair is the Perceiving bipolar scale. *Types* refer to the four-way combinations or interactions of these traits; e.g. ENTP is a type. Types are usually associated with Jungian personality type theory. *Temperaments* are specific two-way combinations of traits and are usually associated with temperament theory.

3. Temperament theory does not claim to explain all – or even a majority – of the aspects of personality but it does claim to make testable predictions about personality. Like any scientific model, it oversimplifies but hopefully not to the point of being useless for experimental purposes.

4. For an in-depth review of this history, see Keirsey (1998).

5. In should be noted that the MBTI is in a similar situation with regard to its foundations. The theory behind the MBTI derives from Jungian psychology, which is primarily a clinical and observation-based psychology, not an experimental one (Spoto 1989). Jung (1971) states that he can give "no a priori reason for selecting" the divisions he used (p. 437). The input-process-output model used in this paper to explain some of the theoretical constructs are from systems theory and are a latter addition. The experimental support for Jungian theory derives largely from the results obtained from employing the MBTI. Similarly, the experimental support for temperament theory should be based on the effectiveness of the KTS.

6. Jung (1971) explicitly developed a model with 8 types. As discussed further in the Jungian Personality Type Theory section, the model was expanded to 16 types. Most current Jungians accept this expanded model as consistent with ideas implicit in Jung's psychology. Thus, these 16 types are generally referred to as the 16 Jungian types.

7. There is some disagreement among the experts on this point. The orthodox view is that very little change in the personality-type occurs over time. Myers et al. (1998: 164) reports evidence that an informal study conducted over a 50-year period found that 54% of the participants changed either none of their four preferences or only one. The probability of this occurring by chance is 6.25% (Myers et al., 1998). Some, however, hold that radical changes in the individual's personality type can occur during middle age (Mid-life Crisis) (see Pascal, 1992). Another interesting point here is Jung's belief that a perfectly formed individual would possess all 6 traits equally and in balance (Jung, 1971). However, he thought it extremely unlikely that this ever occurs, especially since the personality-structure is strongly set during early childhood. Furthermore, an individual with this ideal personality would probably not have a "personality" or "self" in the traditional sense. The homogeneity of traits would result in an individual without the kinds of differentiation that are used to identify different types of persons.

8. This is a good illustration of Jung's de-emphasis on the unconscious in contrast to Freud.

9. Since the MBTI and KTS are basically measuring the same eight traits or variables (Extraversion, Introversion, Intuition, Sensing, Thinking, Feeling, Judging and Perceiving), the two-way combinations derived from the MBTI should also indicate the corresponding temperaments. There are some slight differences in how the KTS and MBTI define and measure the eight traits, as to be expected since the two instruments

use different sets of questions. Studies between the two instruments in regard to the eight trait variables indicate correlations in the range of 70% (Berens, 1996; Keirsey website, 2002). Thus, at the level of the individual eight variables, the two instruments may be used interchangeably, with some precautions.

10. For the website version, the profile analysis of the four temperaments is free, but the 16 type version using four letters similar to the MBTI is available only if purchased.

11. For examples of personality type-based research in non-accounting areas, see Gardner & Martinko, 1996; Hammer, 1996; and Myers et al., 1998.

12. There are other reasons to consider using the KTS instead of the MBTI. These are primarily logistical and are therefore of secondary importance.

(a) The KTS is shorter. It requires less time, which is often an important consideration when trying to use a psychometric instrument in conjunction with another experimental task.
(b) The online version of the KTS is immediately available. It can be administered in any lab with an Internet connection. The scoring is also done online, relieving the researcher of this task, which may introduce errors. The MBTI must be purchased, installed, have a license agreement, etc.
(c) The researcher does not need to be "qualified" to administer the KTS. MBTI requires researchers to have a certain degree of training.
(d) While the online KTS (KTS website, 2002) is not completely free, it is less expensive than the online MBTI. The KTS website (2002) provides the temperament results free, but full scoring of the eight separate traits (i.e. the four letter version of 16 types) must be purchased. Paper versions of the KTS are also less expensive than those of the MBTI. Furthermore, *Please Understand Me II* (1998) includes a paper copy of the KTS without restricting its use. If the two-way interactions of the four temperaments are all that an individual researcher or educator is interested in (rather that one of the 16 four letter types), then the KTS website (2002) provides this service at no charge.

REFERENCES

Baron, J. (1994). *Thinking and Deciding* (2nd ed.) New York, NY: Cambridge University Press.

Berens, L. (1996). Type and temperament. *Bulletin of Psychological Type, 19*, 8–9.

Berens, L. (1998). *Understanding Yourself and Others: An Introduction to Temperament.* Huntington Beach, CA: Telos Publications.

Berens, L. (1985). A Comparison of Jungian Function Theory and Keirseyan Temperament Theory in the Use of the Myers-Briggs Type Indicator. Unpublished doctoral dissertation, United States International University.

Chenhall, R., & Morris, D. (1991). The effect of cognitive style and sponsorship bias on the treatment of opportunity costs in resource allocation decisions. *Accounting, Organizations and Society, 16*, 27–46.

DiTiberio, J. K. (1996). Education, learning styles, and cognitive styles. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 123–166). Palo Alto, CA: Consulting Psychologists Press.

Feist, J., & Feist, G. J. (1998). *Theories of Personality* (4th ed.). Boston, MA: McGraw-Hill.

Flanagan, O. (1991). *The Science of the Mind* (2nd ed.). Cambridge, MA: MIT Press.

Frisbie, G. R. (1988). Cognitive styles: An alternative to Keirsey's temperaments. *Journal of Psychological Type, 16*, 13–21.

Gardner, W. L., & Martinko, M. J. (1996). Using the Myers-Briggs Type Indicator to study managers: A literature review and research agenda. *Journal of Management, 22*, 45–83.

Gul, F. A., & Fong, S. C. C. (1993). Predicting success for introductory accounting students: Some further Hong Kong evidence. *Accounting Education, 2*, 33–42.

Haley, U. C. V. (1997). The MBTI and decision-making styles: Identifying and managing cognitive traits in strategic decision-making. In: C. Fitzgerald & L. K. Kirby (Eds), *Developing Leaders: Research and Applications in Psychological Type and Leadership Development* (pp. 187–223). Palo Alto, CA: Davies-Black.

Hall, C. S., & Nordby, V. J. (1973). *A Primer of Jungian Psychology*. New York, NY: New American Library, Inc.

Hammer, A. L. (1996). Career management and counseling. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 31–53). Palo Alto, CA: Consulting Psychologists Press.

Hammer, A. L., & Huszczo, G. E. (1996). Teams. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 81–103). Palo Alto, CA: Consulting Psychologists Press.

Harvey, R. J. (1996). Reliability and validity. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 5–29). Palo Alto, CA: Consulting Psychologists Press.

Hergenhahn, B. R., & Olson, M. H. (1999). *An Introduction to Theories of Personality* (5th ed.). Upper Saddle River, NJ: Prentice Hall, Inc.

Hobby, S. A., Miller, D. I., Stone, J. A., & Carskadon, T. G. (1987). An empirical test of differing theoretical positions of Myers and Keirsey concerning type similarity. *Journal of Psychological Type, 13*, 56–60.

Hozik, J., & Wright, J. W., Jr. (1996). A cross-cultural investigation of personality traits among Arab and American business students. *Social Behavior and Personality, 24*, 221–230.

Jung, C. G. (1971[1921]). *Psychological Types*. Princeton, NJ: Princeton University Press.

Kahneman, D., Slovic, P., & Tversky, A. (Eds) (1982). *Judgment under Uncertainty: Heuristics and Biases*. New York, NY: Cambridge University Press.

Keirsey, D. (1998). *Please Understand Me II: Temperament, Character, Intelligence*. Del Mar, CA: Prometheus Nemesis Book Company.

Keirsey, D. (1991). *Portraits of Temperament*. Del Mar, CA: Prometheus Nemesis Book Company.

Keirsey website (2002). The temperaments. URL: http://www.keirsey.com

Kelly, K. R., & Jugovic, H. (2001). Concurrent validity of the online version of the Keirsey Temperament Sorter II. *Journal of Career Assessment, 9* (1, Winter), 49–59.

Kirby, L. K., & Barger, N. (1996). Multicultural applications. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 167–196). Palo Alto, CA: Consulting Psychologists Press.

Laribee, S. (1994). The psychological types of college accounting students. *Journal of Psychological Type, 28*, 37–42.

McCarley, N. G., & Carskadon, T. G. (1986). The perceived accuracy of elements of the 16 type descriptions of Myers and Keirsey among men and women: Which elements are most accurate, should the type descriptions be different for men and women, and do the type descriptions stereotype sensing types? *Journal of Psychological Type, 11*, 2–29.

Meier, C. A., & Wozny, N. A. (1978). An empirical study of Jungian typology. *Journal of Analytical Psychology, 23*, 3–15.

Myers, I. B., & McCaulley, M. (1985). *Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.

Myers, I. B., & McCaulley, M. Quenk, N. L., & Hammer, A. L. (1998). *MBTI Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Myers, I. B., & Myers, P. B. (1995). *Gifts Differing: Understanding Personality Type*. Palo Alto, CA: Davies-Black Publishing.

Nourayi, M. M., & Cherry, A. C. (1993). Accounting students' performance and personality types. *Journal of Education for Business*, (November/December), 111–115.

Nunnally, J. M., & Bernstein, I. (1994). *Psychometric Theory* (3rd ed.). New York, NY: McGraw-Hill Book Company.

Oswick, C., & Barber, P. (1998). Personality type and performance in an introductory level accounting course: A research note. *Accounting Education, 7*, 249–254.

Pascal, E. (1992). *Jung to Live By*. New York, NY: Warner Books, Inc.

Quenk, N. L., & Quenk, A. T. (1996). Counseling and psychotherapy. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 105–122). Palo Alto, CA: Consulting Psychologists Press.

Quinn, M. T., Lewis, R. J., & Fisher, K. L. (1992). A cross-correlation of the Myers-Briggs and Keirsey instruments. *Journal of College Student Development, 33*, 279–280.

Roback, A. A. (1973). *The Psychology of Character*. New York, NY: Arno Press.

Rosenak, C. M., & Shontz, F. C. (1988). Jungian Q-Sorts: Demonstrating construct validity for psychological type and the MBTI. *Journal of Psychological Type, 15*, 33–45.

Ruhl, D. L., & Rodgers, R. F. (1992). The perceived accuracy of the 16 type descriptions of Myers and Keirsey: A replication of McCarley and Carskadon. *Journal of Psychological Type, 23*, 22–26.

Satava, D. (1997). Extroverts or introverts: Who supervises the most CPA staff members? *Journal of Psychological Type, 43*, 40–43.

Shelton, J. (1996). Health, stress, and coping. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 197–215). Palo Alto, CA: Consulting Psychologists Press.

Sipser, M. (1997). *Introduction to the Theory of Computation*. Boston, MA: PWS Publishing Company.

Spoto, A. (1989). *Jung's Typology in Perspective*. Boston, MA: SIGO Press.

Swanger, N. A. (1998). Quick service chain restaurant managers: Temperament and profitability. Unpublished doctoral dissertation, University of Idaho.

Tucker, I. F., & Gillespie, B. V. (1993). Correlations among three measures of personality types. *Perceptual and Motor Skills, 77*, 650.

Tzeng, O. C. S., Ware, R., & Chen, J. (1989). Measurement and utility of continuous unipolar ratings for the Myers-Briggs Type Indicator. *Journal of Personality Assessment, 53*, 727–738.

Vassen, E., Baker, C., & Hayes, R. (1993). Cognitive styles of experienced auditors in the Netherlands. *British Accounting Review, 25*, 367–382.

Walck, C. L. (1996). Management and leadership. In A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 55–79). Palo Alto, CA: Consulting Psychologists Press.

Ware, R., & Yokomoto, C. (1985). Perceived accuracy of Myers-Briggs Type Indicator descriptions using Keirsey profiles. *Journal of Psychological Type, 10*, 27–31.

Waskel, S. A. (1995). Temperament types: Midlife death concerns, demographics and intensity of crisis. *The Journal of Psychology, 129*, 221–233.

Wheeler, P. R. (2001). The Myers-Briggs Type Indicator and applications to accounting education and research. *Issues in Accounting Education, 16*, 125–150.

Wolk, C., & Nikolai, L. A. (1997). Personality types of accounting students and faculty: Comparisons and implications. *Journal of Accounting Education, 15*, 1–17.

THE KEIRSEY TEMPERAMENT SORTER: INVESTIGATING THE IMPACT OF PERSONALITY TRAITS IN ACCOUNTING

Patrick Wheeler, Carol Jessup and Michele Martinez

ABSTRACT

The Keirsey Temperament Sorter (KTS) is a psychometric instrument that can be useful to researchers interested in investigating the impact of personality traits in accounting practice and education. The KTS may be used to investigate: (a) the nature of personality of accounting practitioners, faculty and students; and (b) how personality traits affect the performance of accounting practitioners, faculty and students. This paper provides KTS users with the empirical and conceptual information necessary to conduct research in these areas. The psychometric properties and underlying theory of the KTS are examined, as are its limitations. The paper also compares the KTS to the Myers-Briggs Type Indicator (MBTI), another widely used psychometric instrument. The two instruments are compared in order to determine the relative advantages of each for accounting research.

Advances in Accounting Behavioral Research, Volume 5, pages 247–277.

ISBN: 0-7623-0953-9

INTRODUCTION

Instruments for determining psychological characteristics and personality traits have been used in accounting research, accounting education and the accounting profession for several decades (Wheeler, 2001). Some of these instruments derive from personality theories, a cluster of theories comprising one of the established branches of psychology (Feist & Feist, 1998; Hergenhahn & Olson, 1999). For example, the Myers-Briggs Type Indicator (MBTI) has been used in industry and psychology for over 40 years and in accounting research for over 20 years (Myers et al., 1998; Wheeler, 2001). This instrument is based on C. G. Jung's theory of personality type.

Another more recent personality trait instrument is the Keirsey Temperament Sorter (KTS). The KTS is based on temperament theory, a personality theory that is indirectly related to Jungian personality type theory. The KTS, like the MBTI, has been used by businesses and in education and has recently begun appearing in published academic accounting research (e.g. Gul & Fong, 1993). The purpose of this paper is to examine the theoretical and practical aspects of the KTS, especially in comparison to the MBTI. Accordingly, both temperament theory and Jung's personality type theory will be discussed. The research and educational applications of the KTS will also be examined in detail.

From a research perspective several issues surround the use of the KTS as an instrument for conducting behavioral research. These issues may be categorized as: (1) the theoretical underpinning of the instrument; (2) the psychometric properties of the instrument; and (3) the areas of application in which the instrument may be used. The first category of issues is concerned with the constructs that the KTS purports to capture and measure. While the logical and philosophical integrity of a theory is important, the ability of an instrument to measure variables derived from theory is equally important, particularly from an empirical perspective, since this capability directly impacts how the theory is applied. There are numerous psychological instruments, none of which can reasonably claim to capture the whole spectrum of personality traits; therefore, the experimenter must carefully choose the instrument to match the specific needs of the experiment being conducted.

The second category, psychometric properties, is concerned with reliability and validity of the instrument. Reliability testing examines the consistency of the instrument over several dimensions; e.g. time, populations and different versions of the instrument. Validity testing is concerned with the theoretical soundness of the instrument.

The third category of issues, areas of application, examines the types of questions that can be investigated by the instrument. First, as noted above, the

KTS is not suitable for investigating all psychological or even personality research questions. For example, according to theory, the personality traits in temperament theory are predicted to be highly stable over time, if not immutable. Thus, researchers should not use the KTS to capture before-and-after treatment effects. Second, the KTS differs from Jungian-based instruments, in particular the MBTI. The KTS and the MBTI should not be applied in the same manner. Both the KTS and the MBTI purport to measure the four Jungian personality bipolar scales: Extraversion-Introversion; Sensing-Intuition; Thinking-Feeling; Judging-Perceiving, respectively.[1] However, while the MBTI is especially concerned with the four-way interaction of the bipolar scales as the basis for the personality *types*, the KTS is concerned with two-way interactions as the basis for personality *temperaments*.[2]

In this paper, these three categories of issues will be discussed, with emphasis on their significance to accounting. The KTS is an instrument with much potential for furthering accounting research. To use this instrument properly, understanding its underlying theory and its psychometric properties is necessary. Understanding the relationship between the KTS and the MBTI is also helpful, both how they overlap and differ. Once these areas are examined, the wide range of potential research topics made possible by the KTS can be explored.

The remainder of this paper will first discuss temperament theory, the personality theory underlying the KTS, followed by an examination of Jung's theory of personality types. In the context of this paper, there are two reasons for examining Jungian theory. First, although temperament theory is distinct from Jungian theory, the two are closely, if indirectly, related. This complex relationship allows the KTS, as an instrument, to use variables developed by Jungian theory as the specific personality traits it attempts to measure. Thus, both the KTS and MBTI measure the same personality-trait variables, although they interpret them and their interactions differently. The second reason for discussing Jungian theory is the need to compare the psychometric properties and experimental uses of the two instruments. Following the discussion of Jungian theory, the KTS as a psychometric instrument is examined in detail, including its validity and reliability. The KTS is compared to the MBTI in this section. Next, research opportunities using personality theory instruments are delineated, with particular attention paid to the KTS. Finally, there are concluding comments, including the limitations of using the KTS and other personality trait instruments for accounting research.

TEMPERAMENT THEORY

Temperament theory postulates four fundamental temperaments as the source of many observable human characteristics and predictable behaviors.[3] A

temperament may be defined as a cluster of distinct and interrelated personality characteristics or traits. While an individual will have a mix of some or all of the four temperaments, one temperament will clearly dominate the personality according to temperament theory (Keirsey, 1998; Berens, 1996, 1998). This mixture of the temperaments, including the dominant temperament, is theorized to be inherited. Therefore, the temperament structure (i.e. mixture and dominance) is a highly stable aspect of the personality. Although the temperament undergoes development as the individual matures, temperament theory postulates that the basic structure stays relatively unchanged.

Temperament theory has a history dating back to at least Hippocrates' theory of the four humors in the 5th century BC. A review of temperament theory indicates that the theory has influenced a wide range of disciplines throughout the centuries, including modern philosophy, psychology and medical science (Roback, 1973). Temperament theory influenced Jung's formulation of his particular personality type theory (Jung, 1971; Keirsey, 1998). This link is especially relevant because this indirect relationship allows the KTS to utilize personality variables defined in Jungian psychology, although they are not strictly temperament theory constructs. The four temperaments have received various labels over time.[4] Because our focus in this paper is on applying temperament theory in research, especially via the KTS, we will adopt Keirsey's terminology. Keirsey labeled the four temperaments as: (1) Artisan, (2) Guardian, (3) Idealist, and (4) Rational. In the next section, an overview of the major defining characteristics of the temperaments is presented. These definitions are intended to present the accounting researcher with an overview of the kinds of characteristics that the KTS can provide for experimental purposes; they are not comprehensive. For detailed presentations see Keirsey (1998) and the Keirsey website (2002).

The Four Temperaments

Individuals with ***dominant Rational temperaments*** tend to be abstract and theory-oriented in their cognitive aspects, yet pragmatic and utilitarian in applying their ideas. They can be highly accomplished in strategy and in depth analysis. In terms of organizations, they are strong at marshalling resources and planning; in terms of applications, they are skilled at inventing and configuring. They take pride in being competent in their actions. They respect themselves for their ability to be autonomous and feel confidence because of their strong willpower. They tend to view the search for knowledge as a defining aspect of their personality. Thus, they trust in reason and are high achievers. Although an unusually small part of the population (5% to 7% according to the Keirsey website (2002)), they are frequently found in the sciences, technology and systems work.

Individuals with ***dominant Idealist temperaments*** tend to be cognitively abstract and theoretical, like those with Rational temperaments. However, the Idealist approach to applying and implementing their ideas is more group-oriented and cooperative than that of Rationals. Idealists tend to be skillful at diplomacy and integrating the contributions of others into their own work. Thus, they make excellent mentors and are often found among teachers and counselors. They also are outstanding leaders and high-level managers. They value highly social development and view interpersonal integration as a primary defining characteristic of their personality. Accordingly, they take pride in being ethical, empathetic and respectful of others. They tend to view the search for a unique identity, along with meaningful relationships, of primary importance. They are only slightly more frequent than Rationals, making up between 8% and 10% of the population (Keirsey website, 2002).

Individuals with ***dominant Artisan temperaments*** tend to be concrete, down-to-earth thinkers, with practical and utilitarian approaches to implementing goals. They can be extremely skillful in day-to-day operations and tactics. Thus, they are proficient at promoting and expediting plans. They also are skillful at composing and improvising. To Artisans, efficiency of action and elegance of solution tend to be as important as effectiveness and accuracy. They frequently measure their self-esteem in terms of gracefulness in action, degree of daring, and adaptability to change. According to the Keirsey website (2002), they comprise between 35% and 40% of the population. They gravitate towards careers in the arts and crafts, vocations requiring techniques, and operations work.

Individuals with ***dominant Guardian temperaments*** are also concrete, down-to-earth thinkers, similar to those with Artisan temperaments. However, Guardians tend to be more group-oriented in implementing and applying their ideas than are Artisans. They are thus highly proficient in logistics, with particular skills in supplying, conserving and protecting. They also display expertise in administration, supervising, auditing and inspecting. They pride themselves in being reliable and dependable in action, and valuing respectability highly. Security is central to their self-image, whether providing security, in the sense of being responsible for the goods of others, or receiving security, in the sense of being trusted and accepted by others. They are frequently found in careers involving material work and logistics, commerce and regulations. According to the Keirsey website (2002), they are the most numerous of the four temperaments, comprising from 40% to 45% of the population.

Keirsey, who developed the KTS, further defines the four temperaments to include concepts from linguistics and anthropology (Keirsey, 1991, 1998). From linguistics, Keirsey distinguishes those who prefer abstract communications

from those who prefer concrete communications. This aspect of the temperaments is measured by observing the types of words used. From anthropology's emphasis on the human reliance on tools, Keirsey distinguishes cooperative tool users from utilitarian or pragmatic tool users. This aspect is measured by observing the use of tools ranging from a simple one, such as a hammer, to the complex, such as a computer.

The intersection of the communications and tools aspects gives rise to a 2×2 four-cell matrix. Keirsey (1991, 1998) identified the resulting four combinations with the four temperaments as follows:

- Concrete communicator, pragmatic tool-user = Artisan
- Concrete communicator, cooperative tool-user = Guardian
- Abstract communicator, pragmatic tool-user = Rational
- Abstract communicator, cooperative tool-user = Idealists

Language and tool usage are generally accepted as being activities that distinguish human behavior from other types of behavior. These are activities that all people engage in on a routine basis, yet complex level. These aspects of human behavior can be reasonably used to distinguish between personalities. Thus, by using these two fundamental aspects of human behavior, Keirsey provides a scientific framework for understanding and justifying temperament theory's division of personality into four kinds. Previously, this four-way division of personalities was primarily based on psychological observation and philosophical speculation. By adding results from the sciences of anthropology and linguistics, Keirsey places temperament theory and the KTS on a more solid foundation, especially important for scientific researchers.[5]

Keirsey provides the experimental user of the KTS another invaluable service by incorporating personality constructs from the Jungian/MBTI personality type theory into the KTS. To examine this complex relationship between the temperament/KTS theory and Jungian/MBTI theory, Jungian personality type theory is discussed in the next section. This is followed by a section on the relationship between these two personality theories. The subsequent discussion then provides an in depth examination of the properties of the KTS as an instrument for measuring personality characteristics.

JUNGIAN PERSONALITY TYPE THEORY

The relationship between temperament theory and Jungian personality type theory is complex. There is general acceptance that the two theories are distinct yet with considerable overlap (Keirsey, 1998; Myers et al., 1998). Temperament theory is not, however, a derivative of Jungian theory, a mistake easily made when

one considers the similarities between the KTS and MBTI. Furthermore, both theories fall within the set of psychology theories known as personality theories (Feist & Feist, 1998; Hergenhahn & Olson, 1999). Beyond these commonalities, however, the two theories diverge. Temperament theory uses a model based on a four-way division of kinds of personalities (i.e. the four temperaments); Jungian theory uses a model of 16 types.[6] It is necessary, nevertheless, to discuss Jungian theory in regard to the KTS because the personality-trait variables used in the KTS (viz., Extraversion, Introversion, Sensing, Intuition, Thinking, Feeling, Judging and Perceiving) are directly derived from Jungian theory.

C. J. Jung's theory of personality types was developed in his 1921 work, *Psychological Types,* in response to his break with Freud over the latter's emphasis on the individual's unconscious (Jung, 1971; Hall & Nordby, 1973). Jung's approach has a strong information processing element, focusing on decision-making and the effect of personality on understanding of the world. Because of this orientation, Jungian theory shares many concerns and features with the currently dominant cognitive science model of psychology (Flanagan, 1991). Both approaches see information processing as a primary process or function of the mind. Cognitive psychology is based on a computational model of information processing developed from mathematical theory (Sipser, 1997). This approach, which also underlies computer science, tends to de-emphasize the role of personality in information processing. Research into the heuristics and biases used by humans, however, suggests that a predominantly computational model does not adequately explain the processes involved in human information processing (Kahneman et al., 1982; Baron, 1994). Jung's theory, with its inclusion of both information processing and personality traits, provides theoretical cohesion for the scattered results of heuristics and biases research and can supplement the cognitive science model of the mind with a needed synthesis of information processing and personality.

Jung's (1971) original personality type theory posits six personality traits which an individual uses to process information from the world. These six traits are grouped into three bipolar scales (i.e. complementary, dichotomous pairs). Four of the six traits derive from two fundamental mental processes (*Judging* and *Perceiving*) while the remaining two traits reflect an overall attitude toward the world (*Extraversion* and *Introversion).* Accordingly, the two fundamental mental processes (Judging and Perceiving) result in four mental functions – *Sensing, Intuition, Thinking* and *Feeling.* Individuals constantly use these four basic mental functions to perceive the world and construct their world-views. Sensing and Intuition are perceiving functions and deal with the type of inputs used for mental processing. Thinking and Feeling are Judging functions and transform the Sensing or Intuition inputs into each individual's unique world-view.

Furthermore, all individuals have relative interests in, or attitudes toward, the inner and outer worlds that affect their mental processes – *Extraversion* and *Introversion.*

Each individual is a mix of the six personality traits. However, Jung's theory postulates that, from each of the three pairs (Sensing or Intuition, Thinking or Feeling, and Extraversion or Introversion), one trait will be dominant or preferred. The resulting set of three *preferred* traits determines the personality type of the individual. Furthermore, one of the four mental functions will be dominant (Sensing, Intuition, Thinking or Feeling) leading to a dominant mental process (Perceiving or Judging). That is, a person will be dominantly a Judging-preference individual (Thinking or Feeling) or a Perceiving-preference individual (Sensing or Intuition). The three non-preferred traits are still present but have secondary roles in the nature of the personality. (By analogy, people have a preference for using either the right hand or left hand, and are therefore described as right-handed or left-handed. Nevertheless, everyone has some capability for using the non-preferred hand.) The resulting combination of preferred traits yields eight personality types as shown in Table 1.

Table 1. Jung's Eight Personality Types.

Three Trait Designation	Dominant Mental Function	Eight Personality Types
EST ENT	Judging (T or F)	Extraverted, dominant Thinking: The Extraverted Thinking type
EST ESF	Perceiving (S or N)	Extraverted, dominant Sensing: The Extraverted Sensing type
ESF ENF	Judging (T or F)	Extraverted, dominant Feeling: The Extraverted Feeling type
ENT ENF	Perceiving (S or N)	Extraverted, dominant Intuitive: The Extraverted Intuitive type
IST INT	Judging (T or F)	Introverted, dominant Thinking: The Introverted Thinking type
IST ISF	Perceiving (S or N)	Introverted, dominant Sensing: The Introverted Sensing type
ISF INF	Judging (T or F)	Introverted, dominant Feeling: The Introverted Feeling type
INT INF	Perceiving (S or N)	Introverted, dominant Intuitive: The Introverted Intuitive type

Notes: The following standard abbreviations apply:
E: Extraversion; F: Feeling; I: Introversion; N: Intuition; S: Sensing; T: Thinking.

When interpreting these personality traits, Jungian theory stipulates that no value judgments be made. This stipulation has two notable implications. First, neither aspect of a bipolar pair of traits is better or worse than the other. For example, being Extraverted is not better than being Introverted. Second, the fact that a trait is preferred (or non-preferred) does not imply how good someone is at using that trait. One may prefer to use the Extraverted trait yet be inept at Extraverted behavior. Conversely, one may have Introversion as a non-preferred trait, yet be very comfortable at acting Introverted. Jungian theory also states that the preferred traits are fixed in the individual from birth or at a very early age, with little subsequent change occurring (Pascal 1992).[7]

Mental Functions

A strong information processing orientation in Jung's theory is apparent in the mental functions. In information systems terms, the mental functions are inputting and processing in nature. However, the theory contains an equally strong behavioral aspect. Perceiving and Judging influence observable behavior. Perceiving, for example, determines which information a person actively gathers as relevant (preferred) in a situation. This information includes both internal and external types of data; i.e. things, people, events, feelings and ideas. Judging takes the information input from the preceding perceptions, and subsequently processes, organizes and transforms it into conclusions. These conclusions range from solutions to particular problems to the development of an integrated, holistic world-view. Note however, that people with different personality types may arrive at similar conclusions and behaviors. This is especially true for scenarios with specific, objectively definable and well-structured problems. But even in these scenarios, the data collection and processing may be markedly different among varying personality types.

Sensing and Intuition are theorized to be two styles of Perceiving, which may be respectively characterized as a preference for seeing either "the trees" or "the forest." These are input functions that describe sources of data to be subsequently processed by the judgment functions. Sensing-preference individuals prefer direct or objective perceptions made through the basic senses; i.e. hearing, sight, etc. Individuals oriented toward Sensing are inclined to center on immediate experience because the senses gather data from events presently happening. They tend to be practical, matter-of-fact, realistic, observant, and pragmatic, and have a good memory of facts.

Intuition-preference individuals prefer subjective perceptions of possibilities, structures, and meanings over sense-based perceptions. They observe associations among objects, experiences, ideas and entities. Individuals who prefer this

function tend to see the overall picture and are imaginative, speculative, abstract and inventive. Such perceptions are often made through insight. Intuitive perception includes concepts and relationships beyond the capabilities of the senses. Intuition should not be equated with vague feelings, since it can be rigorously logical.

Thinking and Feeling are the two types of functions from the Judging bipolar scale. These are processing functions that transform the inputs from the perception functions. Thinking-preference individuals prefer connecting thoughts and experiences together logically. They tend to be analytical and objective, and focus on causal associations. They are especially concerned with principles such as fairness and justice, which may make them seem impersonal when dealing with others. Individuals who prefer Thinking are also frequently systematic problem solvers.

Feeling-preference individuals tend to rely on values when involved in decision-making, either personal or group. Thus, they tend to develop characteristics such as a concern for the human side of problems, sympathy and compassion. A general understanding of people may make them seem to be overly tender hearted. Those with this approach to making judgments are usually more attuned than Thinking-preference individuals to the desires, values and needs of others.

While the four functions may be divided according to whether they are perception-input or judgment-processing, they also compete with each other to provide the mind with a preferred objective or viewpoint in understanding the world. Each function provides a different objective to conscious mental activity. Sensing offers to the mind a full experience of the immediate and real sensory world. Intuition offers a deep comprehension of the structure and possibilities of the world, both imaginative and sensory. Thinking offers the mind a logically and rationally ordered world. Feeling offers a rational world also, but one built around an orderly arrangement of subjective values. These conflicting goals interact with each other so that only two will be primary in the individual, one from each of the two mental bipolar scales. The relative importance of these two preferred mental functions is determined by their relationship with the individual's attitude toward the world.

Attitudes Towards Internal and External Aspects of the World

According to Jung's theory, Extraversion and Introversion are fundamental attitudes that describe the individual's approach toward the world in its internal and external aspects. Jung believed that this distinction was the most fundamental in the personality of the individual (Jung, 1971). While a healthy life requires a mix of both attitudes, one clearly dominates in the personality.

Extraversion is an attitude wherein the individual's focus is on people and objects of the external world. Extraverts are keenly sensitive toward the environment. For Extraverts, external objects – personal and impersonal – are sources of energy and direction. They focus their particular Perceiving-functions (Sensing-Intuition) and Judging-functions (Thinking-Feeling) on things in the outside environment. Extraverts receive the energy for these mental processes from experiences involving external activities and events. Extraverts tend to be action-oriented, expressive, and outgoing. They excel at oral communication and have learning-styles that are group-based and action-based.

Introversion is an attitude wherein the individual's attention is directed primarily to the inner environment of the mind and those things that comprise this subjective world of thoughts and feelings. They get their energy from the activities of the mind; i.e. thinking, feeling, and reflecting; and they direct their energy and actions toward this realm. They focus their particular Perceiving-functions (Sensing-Intuition) and Judging-functions (Thinking-Feeling) on things in the internal environment. They put their energy into concepts and ideas, committed to making them clear and accurate. Those who prefer Introversion tend to be contemplative and detached from the external world. They enjoy and need solitude and privacy. This is when they re-energize. They prefer written communication and have individual learning-styles.

Developing a Fourth Bipolar Scale: Orientation towards the External World

Jung theorized that between the two preferred functions from the two mental bipolar scales, one function would be dominant. This dominance defines whether a person is primarily interested in data collection or data processing. When the Perceiving bipolar scale (Intuition-Sensing) is dominant, one is more interested in gathering information; when the Judging bipolar scale (Thinking-Feeling) dominates, one is more concerned about solving problems and arriving at conclusions. A review of Table 1 reveals that the non-dominant yet preferred mental function was essentially "dropped" from Jung's description of the personality (Myers & Myers 1995). For example, an EST who is dominantly a judging-preference individual is described as "an extraverted thinking type"; i.e. the preferred Sensing function is not included in the description. Jung believed that this "dropped" mental function, referred to as the auxiliary function, was the least differentiating of the functions in the personality, primarily because it functioned unconsciously.[8]

I. B. Myers modified Jung's theory further by developing a fourth personality bipolar scale. (See Myers & Myers, 1995 for a discussion of the history of this development.) This new scale was introduced in response to mixed

results from empirical work attempting to test Jung's theory (Meier & Wozny, 1978; Rosenak & Shontz, 1988). These results indicated that Jung's eight personality types were not sufficient to classify the data concerning personality characteristics. The new bipolar scale was introduced to clarify ambiguities about the interrelationship or interaction effect among the preferred functions in the mental bipolar scales. Specifically, it was developed to indicate whether an individual shows an overall preference for the preferred function in the Judging bipolar scale (Thinking-Feeling) or the preferred function in the Perceiving bipolar scale (Sensing-Intuition) *when dealing with the external world.* One result of introducing this fourth bipolar scale was to make the auxiliary function an important aspect of the personality type.

This new bipolar scale from I. B Myers, often referred to as an attitude or orientation to the external world, consists of two traits, *Judging* and *Perceiving.* The Judging and Perceiving traits in the fourth bipolar scale may initially appear to be present in Jung's theory. To some extent this is true. As discussed above, the Judging mental function contains Feeling and Thinking functions, and the Perceiving mental function contains Sensing and Intuition functions. However, the new (fourth) bipolar scale is an attitude that indicates one's preference for either the Judging or Perceiving mental function in relation to the external world, regardless of whether one is otherwise an Extravert or Introvert. That is, what the fourth bipolar scale is trying to ascertain is whether the individual prefers using one of the Judging mental functions (Thinking or Feeling) or one of the Perceiving mental functions (Sensing or Intuition) when dealing with the outside world. The terminology overlaps and may seem confusing (see note 2).

Judging-preference individuals are those who prefer to use the Judging mental function when dealing with the outer world. Judging-preference individuals prefer planning and organizing activities. They like to see problems solved to completion. They especially desire closure in their mental activities and are, therefore, thorough, conscientious, methodical and decisive. Perceiving-preference individuals are those who prefer to use the Perceiving bipolar scale when dealing with the external world. They prefer openness in their mental activities and are concerned that all possible information is gathered. They like to keep the problem open as long as possible. This tends to make them spontaneous, curious and adaptive decision makers. From an information processing perspective, Judging-preference individuals may be characterized as individuals who like to perceive the output as completed and closed to further revision. Perceiving-preference individuals prefer to see the output as incomplete and open to additional processing.

In summary, Jungian personality type theory, postulates eight personality traits from four bipolar scales: Extraversion-Introversion; Sensing-Intuition;

Thinking-Feeling; Judging-Perceiving. All eight traits are present in each person and can be treated as continuous variables. This continuous approach is not, however, the usual interpretation attached to the presence of the traits in individuals. The traits are generally treated as dichotomous and exclusive. In an individual, one trait from each of the four bipolar scales is preferred. According to Jungian theory, although individuals have all eight traits to some extent, the four preferred traits primarily define the personality. This system of preferences results in 16 possible combinations of the four preferred traits, which are referred to as the 16 Jungian personality types. As will be discussed in the next section, this system of four bipolar scales is used in the KTS (Keirsey, 1998). Table 2 provides general characteristics and occupational tendencies of the 16 Jungian personality types.

The Relationship between Temperament Theory and Jungian Theory

As noted earlier, temperament theory is not an offshoot of Jungian personality type theory. Nor does Jungian theory derive from temperament theory, although the latter predates the former and influenced Jung. Yet, the above excursion into Jungian theory is necessary in order to understand Keirsey's formulation of temperament theory and the Keirsey Temperament Sorter because Keirsey uses the eight Jungian personality traits in his system. His justification for doing so is primarily observation-based, not theory-driven (Keirsey, 1998; also see note 5 on this aspect of personality theories in general). Although Keirsey believes that Jungian theory is "cumbersome and self-contradictory" (Keirsey, 1998, p. 15), he accepts the constructs underlying the MBTI as valid and reliable. According to Keirsey (1998), the individual traits or variables in the MBTI (i.e. Extraversion, Introversion, Sensing, Intuition, Thinking, Feeling, Judging and Perceiving) accurately measure characteristics of the personality, but he disagrees with how MBTI/Jungian theory subsequently combines and interprets these individual variables (discussed in detail later).

KTS temperament theory is based on a systems model, focusing on the configuration of the whole. MBTI Jungian theory, on the other hand, is based on a dynamic parts model, in which the attitudes and functions can be consciously manipulated. Keirsey and Myers disagree in what constitutes similar personalities, primarily because of their different categories for grouping the various personality traits. Myers groups traits according to Jungian type theory, while Keirsey groups traits according to temperament theory. Divergences in Keirsey's and Myers' "similar types" have been empirically tested and are discussed below (Hobby et al., 1987).

Table 2. The 16 Personality Types with General Characteristics and Occupational Tendencies.

	Sensing (S)		Intuition (N)		
	Thinking (T)	**Feeling (F)**	**Feeling (F)**	**Thinking (T)**	
Judging (J)	**ISTJ** Practical, sensible, decisive, logical, detached. *Management and administration.*	**ISFJ** Practical, concrete, cooperative, sensitive. *Education, health care, and religion.*	**INFJ** Insightful, symbolic, idealistic, committed, compassionate. *Religion, counseling, and teaching.*	**INTJ** Insightful, long-range thinkers, clear, rational, detached. *Science, computers, and technical fields.*	**Introversion (I)**
Perceiving (P)	**ISTP** Detached, logical problem solvers, pragmatic, factual. *Skilled trades and technical fields.*	**ISFP** Trusting, kind, sensitive, observant, practical, concrete. *Health care and business.*	**INFP** Sensitive, caring, idealistic, curious, creative, visionary. *Counseling, writing, and arts.*	**INTP** Logical, curious, detached, insightful, contemplative. *Scientific and technical fields.*	
Perceiving (P)	**ESTP** Observant, active, rational problem solvers, assertive. *Marketing, business, and skilled trades.*	**ESFP** Observant, specific, active, sympathetic, idealistic, warm. *Health care and teaching.*	**ENFP** Curious, creative, energetic, friendly, cooperative, warm. *Counseling, religion, and teaching.*	**ENTP** Creative, imaginative, theoretical, analytical, rational, questioning. *Science, management, and technology.*	**Extraversion (E)**
Judging (J)	**ESTJ** Logical, decisive, objectively critical, practical, systematic. *Management and administration.*	**ESFJ** Factual, personable, cooperative, practical, decisive. *Education, health care, and religion.*	**ENFJ** Compassionate, loyal, imaginative, likes variety, supportive. *Arts, religion, and teaching.*	**ENTJ** Analytical, assertive, conceptual thinkers, innovation planners. *Management and leadership.*	

Notes: General personal characteristics are shown in regular font style; occupational tendencies are shown in italics.
Source: Reproduced from Wheeler (2001: 128) with permission.

Keirsey (1998) uses the Jungian-based variables to arrive at operationalized definitions of the four temperaments; i.e. definitions that may be equated with measurements from the KTS. Specifically, he equates the four temperaments with 2-way interactions or combinations of the MBTI/Jungian variables. Thus, an SP (Sensing-Perceiving) combination indicates an Artisan temperament, an SJ (Sensing-Judging) combination indicates a Guardian temperament, an NF (Intuition-Feeling) combination indicates an Idealist temperament, and an NT (Intuition-Thinking) combination indicates a Rational temperament. The KTS, not the MBTI, is used to measure these traits in the subjects for arriving at the temperaments because the KTS operationalizes these traits slightly differently than does the MBTI (i.e. the questions in the two instruments are different). Keirsey is only interested in using the theoretical constructs (i.e. the eight traits) underlying the MBTI, not the MBTI instrument per se. Keirsey contends that these combinations of variables derived from the KTS identify individuals with the characteristics of the corresponding temperament.[9] Table 3 presents the four temperaments in a grid using the eight Jungian/MBTI personality traits.

Some discussion of the relationship between the 16 MBTI types and the four KTS temperaments is required in order to understand how the KTS can be used for research. Temperament theorists, including Keirsey (Keirsey, 1998; Keirsey website, 2002), typically group the 16 Jungian/MBTI types into four groups corresponding to the temperaments. This grouping is done to show the degree of overlap between the two personality theories and to facilitate movement between the various classification schemes. Table 4 provides an example of this approach based on Keirsey's approach (1991, 1998).

As depicted in Table 4, the current names of the Keirsey temperaments are presented with correlation to the MBTI traits. As noted above, both KTS instruments and the MBTI use four bipolar scales and eight traits. There are, therefore, 16 *types* in both Keirsey's and Myers's schemes. However, the two sets of 16 types are not the equivalent. Differences between the two schemes arise, not so much from the measurements made by the respective instrument, but from the descriptions attached to the measured variables. Users mistakenly assume they can interchange Keirsey's type descriptions for the same type letters of the MBTI; while this is frequently done, it is not advised, due to validity concerns pertaining to differing theoretical origins. Keirsey has emphasized that making his 16 types "conform to ancient temperament theory took juggling" (Frisbie, 1988).

In summary, Keirsey uses the eight Jungian/MBTI traits as variables in the KTS and temperament theory because they are effective at measuring detailed characteristics of the personality. He then assembles and interprets the data from these measurements in a manner different than that done by Jungian and MBTI

Table 3. The Four Temperaments with General Characteristics.

	Sensing (S)		Intuition (N)		
	Thinking (T)	Feeling (F)	Feeling (F)	Thinking (T)	
Judging (J)	**SJ** **Guardian temperament** Concrete, cooperative, logistical, dutiful, pessimistic, authoritative, and beneficent. *Administrator, auditor, and judge.*		**NF** **Idealist temperament** Abstract, cooperative, diplomatic, altruistic, credulous, benevolent, and intuitive. *Teacher, ambassador, and doctor.*	**NT** **Rational temperament** Abstract, utilitarian, strategic, pragmatic, skeptical, autonomous, and reasonable. *Inventor, architect, and scientist.*	**Introversion (I)**
Perceiving (P)	**SP** **Artisan temperament** Concrete, utilitarian, tactical, practical, optimistic, audacious, and impulsive. *Salesperson, airline pilot, and entertainer.*				
Perceiving (P)					**Extraversion (E)**
Judging (J)	**SJ** **Guardian temperament** (continued)				

Notes: The Guardian temperament is shown in two sections because of the table format being based on the four bipolar scales. It is, nonetheless, a single temperament like the other three.
General personal characteristics are shown in regular font style; occupational or role tendencies are shown in italics.
Sources: Myers et al. (1998) and Keirsey (1998), with modifications.

Table 4. The Four Keirsey Temperaments and 16 Types, Referencing Correlated MBTI Traits.

	Idealists NF	**Rationals NT**	**Guardians SJ**	**Artisans SP**
Directive role*	**Mentors NFj**	**Organizers NTj**	**Monitors SJt**	**Operators SPt**
Extroverted **(e)**	Teacher eNFj	Field Marshal eNTj	Supervisor eSJt	Promoter eSPt
Introverted **(i)**	Counselor iNFj	Mastermind iNTj	Inspector iSJt	Crafter iSPt
Informative role*	**Advocates NFp**	**Engineers NTp**	**Conservators SJf**	**Players SPf**
Extroverted **(e)**	Champion eNFp	Inventor eNTp	Provider eSJf	Performer eSPf
Introverted **(i)**	Healer iNFp	Architect iNTp	Protector iSJf	Composer iSPf

Notes: Keirsey's terms for the 16 types have evolved. The 16 types presented here are the most recent nomenclature. Myers-Briggs terms have been inserted to indicate correlation, not identity.
* This grouping stems from the kinds of relationships the temperaments are willing to have with others. Typically, roles are seen as primarily directive (i.e. influencing the actions of others) or informative (i.e. providing others with ideas or data) (Keirsey, 1991).
Sources: Keirsey (1998) and Keirsey (1991), with modifications.

theorists. A rough analogy might be one using the inch as a measurement but then aggregating it in groups of ten (as with the metric system) instead of groups of twelve (as with the U.S. customary system).

THE KEIRSEY TEMPERAMENT SORTER

The Keirsey website (2002) makes the claim that the KTS is "the No. 1 online personality test" and that this test is "used by many Fortune 500 companies to test their employees and by major Universities to test their students" (Keirsey website, 2002: Introduction screen). Furthermore, the KTS is used in accounting and business research that has resulted in publications and dissertations (e.g. Gul & Fong, 1993; Hozik & Wright, 1996; Swanger, 1998).

The Keirsey Temperament Sorter (KTS) is a forced-choice test. The forced-choice format consists of a question followed by two responses representing the two traits in one of the four bipolar scales. The initial version was developed and included in Keirsey's 1978 book. The more recent version is referred

to as KTS-II and consists of 70 questions. The KTS-II is available both in a paper format (Keirsey, 1998) and online (Keirsey website, 2002). Both versions of the KTS, similar to the multiple versions of the MBTI, measure the eight Jungian-based traits (i.e. Extraversion, Introversion, Sensing, Intuition, Thinking, Feeling, Judging and Perceiving) and indicate an individual's four preferences from these pairs. However, the resulting personality description from the KTS is based on temperament theory, not Jungian type theory. Temperament theory differs from Jungian type theory underlying the MBTI in that the former places emphasis on the 2-way interactions of certain preferences (Keirsey, 1998; Myers et al., 1998).

Two profiles may be derived from taking and scoring the KTS. The first profile is general to each of four temperaments (SJ, SP, NF, NT); detailed 16 type profiles are also available. Keirsey's book, *Please Understand Me II* includes both temperament and type profiles (Keirsey, 1998).[10] While the KTS uses the same four Jungian-based bipolar scales and accompanying eight traits as the MBTI, the personality profile descriptions provided by the two instruments differ. As will be discussed below, convergent validity studies indicate that KTS and MBTI profiles are, nevertheless, significantly correlated (McCarley & Carskadon, 1986; Hobby et al., 1987).

Validity and Reliability of the Keirsey Temperament Sorter

Issues concerning the validity and reliability of the Keirsey Temperament Sorter are critical because the validity and reliability of an instrument determine its usefulness experimentally. The authors located seven studies examining convergent validity issues surrounding the KTS. Of these, four investigate the relationship between the Keirsey theory and the MBTI (Hobby et al., 1987; McCarley & Carskadon, 1986; Ruhl & Rodgers, 1992; Ware & Yokomoto, 1985). The other three are empirical studies comparing the KTS and the MBTI (Kelly & Jugovic, 2001; Quinn et al., 1992; Tucker & Gillespie, 1993).

To investigate the Keirsey theory, Ware and Yokomoto (1985) first administered the MBTI to undergraduate psychology students to identify their personality traits. Each subject was then given a packet containing several different personality descriptions based on Keirsey's theory, including the subject's type, reversed function type, reversed attitude type, and the opposite type. Next, the subjects were asked to indicate the relative accuracies of the different personality descriptions. Ware and Yokomoto (1985) found that subjects perceived the Keirsey-descriptions of their MBTI-indicated personality type to be more accurate than the other type Keirsey-descriptions (i.e. those not corresponding to the MBTI indicated personality type). Since all of these

descriptions were based on Keirsey's theory, the researchers interpreted this result as supporting the convergent validity of Keirsey's theory, using the MBTI as the validity criterion.

McCarley and Carskadon (1986) used procedures similar to Ware and Yokomoto's (1985) except that they provided the subjects with individual elements from Keirsey's descriptions instead of the whole type descriptions. McCarley and Carskadon (1986) also provided subjects with elements from type descriptions based on the MBTI theory. They found that subjects perceived the elements of Keirsey's descriptions similar to the elements from the MBTI theory.

A replication of McCarley and Carskadon (1986) found similar ratings by participants in overall accuracy of personality descriptors between Keirsey and the MBTI (Ruhl & Rodgers, 1992) that support the earlier study. There was only one significant difference that emerged, Thinking vs. Feeling. Of the sixteen types studied, two of the three types that preferred Keirsey's descriptors were Thinking types, while all three that preferred MBTI were Feeling types.

After determining subject's personality types by administering the MBTI, Hobby et al. (1987) provided each subject with two complete descriptions of the subject's indicated personality type, corresponding to the "most similar" type per Keirsey and per Myers. Note that the type descriptions provided to participants were not of the types as indicated by the MBTI, but were instead what each subject would have perceived as "most similar" to the scored type. The results found for most types no significant difference in how subjects evaluated the relative accuracy of the two type descriptions. In a few types, the MBTI-based descriptions were perceived to be more accurate. However, the results of this study are somewhat ambiguous in that the authors did not use the type descriptions in their original form.

Overall, the studies by Hobby et al. (1987), McCarley and Carskadon (1986), Ruhl and Rodgers (1992) and Ware and Yokomoto (1985) are supportive of the convergent validity of the Keirsey Temperament Sorter. However, two qualifications are necessary to this conclusion. First, these studies investigate only the theory (Keirsey's temperament theory) underlying the instrument (KTS), not the instrument per se. In none of the studies was the KTS used. Second, the studies used the type descriptions from Keirsey's work, not the temperament descriptions; i.e. the two-way interactions were not examined.

The following three studies used the KTS instrument. Kelly and Jugovic (2001) took concurrent measures from undergraduate students from the KTS-II and the MBTI Form G. They found correlations between individual KTS and MBTI traits ranging from a low of 0.60 to a high of 0.78. According to Kelly

and Jugovic (2001, p. 55), these correlations are "moderate to strong" and "indicate that the KTS-II has satisfactory concurrent [i.e. convergent] validity." Using the earlier version of the KTS, Quinn et al. (1992) found correlations between KTS and MBTI traits ranging from 0.54 to 0.74 on a sample of business undergraduates. Tucker and Gillespie (1993), also using the previous version of the KTS, had correlations between the KTS and MBTI ranging from 0.68 to 0.84 on undergraduate psychology students. As with Kelly and Jugovic (2001), these last two studies provide moderate to strong support for the convergent validity of the KTS.

Another study investigated reliability issues of the KTS (Waskel 1995). Waskel (1995) empirically examined the internal consistency of the prior version of the KTS and found alpha coefficients of 0.74 for the Extraversion-Introversion scale, 0.89 for the Sensing-Intuition scale, and 0.87 for the Thinking-Feeling scale and 0.88 for the Judging-Perceiving scale. Alpha coefficients of 0.70 and higher are generally considered acceptable levels of instrument reliability (Nunnally & Bernstein, 1994).

In summary, numerous validity and reliability studies have been conducted on the KTS and its underlying temperament theory. The results consistently provide moderate to strong support for both the instrument and theory. Thus, researchers are justified in assuming that the KTS provides reliable measurements of certain psychological traits and that the descriptions (as derived from temperament theory) of these measured traits have valid content and discriminatory power. One weakness in the validity testing that should be noted is that little has been done using the specific two-way combinations that correspond to the four temperaments.

Comparing the KTS and the MBTI

Like the KTS, the MBTI is a forced-choice test (Myers et al., 1998). It is significantly longer than the KTS, consisting of 93 questions instead of 70, and thus requires more time to administer. The MBTI is copyrighted and can be purchased from Consulting Psychologists Press, Inc. (CPP) in self-scoring, computer mail-in scoring and online versions. The MBTI has evolved since 1942 to its present version (Form M). Form M replaced the prior Form G as of 1998. Both versions are still acceptable for use. CPP requires those administering the MBTI to meet certain psychological testing training and education requirements, which are not required to use the KTS.

The MBTI has undergone extensive reliability and validity testing (Harvey, 1996; Myers et al., 1998; Wheeler, 2001). Tests of internal consistency and temporal stability have consistently provided strong support for the reliability

of the scores from the instrument. Testing of discriminant, convergent and construct validity has not been as unvaryingly supportive of the MBTI as those from reliability studies. The results indicate that the MBTI is measuring aspects of personality in a way usually consistent with Jungian theory but is not capturing the personality in its entirety.

A common criticism of the MBTI and Jungian personality theory is that the former does not capture and the latter does not portray the personality in its entirety. This is a criticism valid for all personality theories, including temperament theory, and probably for any scientific model. Theories and models deliberately oversimplify. The crucial issue is not so much whether the theory captures all of major aspects of the phenomenon under investigation but whether it allows researchers to make discriminatory predictions that are testable. Undoubtedly, a cost-benefit tradeoff is involved; a balance must be struck between increasing understanding and furthering research. In this regard, the KTS and temperament theory are significantly different than the MBTI and Jungian theory, offering the researcher opportunities to make unique predictions. Temperament theory includes the temperaments as constructs not found in Jungian theory. As noted above, little validity testing has been done in this area of the KTS; therefore, this represents an area open to future research.

RESEARCH OPPORTUNITIES USING THE KEIRSEY TEMPERAMENT SORTER

Use of the Keirsey Temperament Sorter for research may be approached from two angles. First, the KTS may be used in a manner similar to the MBTI. The KTS, like the MBTI, results in preference scores for the four bipolar scales. Thus, for example, if one wants to use scores for the two mental bipolar scales to capture cognitive style (as done by Chenhall & Morris 1991 and Vassen et al., 1993), then these scores can be provided by either the KTS or the MBTI. The decision should be based on such considerations as reliability, validity, convenience, availability and cost. Second, as discussed above, unique aspects of temperament theory distinguish it from Jungian personality theory. If the research being undertaken involves temperaments, then the KTS, not the MBTI, should be used.

The KTS can be used to address three general accounting research areas. First, research can be done on college-aged accounting students; e.g. the distribution of indicated personality types and temperaments, and relationships of personality traits to academic performance and choice of major. Second, research can be conducted on teaching styles; e.g. understanding the interaction of teaching styles and learning styles from the respective personality traits

of teachers and students. Third, research in the accounting profession can be done; e.g. the distribution of indicated personality types and temperaments, and relationships of personality traits to career success and job-placement. The use of the KTS within businesses should also be examined since it is widely used for management purposes (Keirsey website, 2002).

Wheeler (2001) gives an extensive review of prior research using the MBTI to investigate the role of personality traits and types in accounting education and the profession. This review is suggestive of numerous specific opportunities for applying personality-based instruments to these three research areas. For example, despite changes in the accounting environment and repeated calls for more diversity in the profession, the distribution of indicated personality traits in the accounting profession has remained remarkably stable.

It appears that much of the stability in personality traits found in the profession results from the college process and environment. One study (Laribee, 1994) indicates that accounting education involves a filtering-out process that decreases the percentages of Extraversion, Intuition, Feeling and Perceiving-preferences among accounting students. Research to explain this filtering is needed. Is it related to teaching-style, although accounting teachers tend to be Intuition-preferenced? (Wolk & Nikolai, 1997). If this filtering is common, discovering ways to decrease it may allow for increased diversity among accounting students. At a time of concern about the future of accounting, this is an area of valuable research.

The learning styles of individuals affect their performance. Personality theory predicts that personality traits affect learning styles (Myers et al., 1998). Accounting researchers should therefore be able to use personality-based instruments to investigate this relationship among accounting students. Also, many other learner characteristics can be investigated using personality-based instruments; e.g. written vs. oral, team vs. individual, concrete vs. theoretical, and structured vs. open-ended problems (Myers et al., 1998).

Studies of the academic performance in undergraduate accounting courses have produced mixed results (Gul & Fong, 1993; Nourayi & Cherry, 1993; Oswick & Barber, 1998). Gul and Fong (1993) and Nourayi and Cherry (1993) found a correlation between personality traits and performance; Oswick and Barber (1998) did not. Further research may be conducted to examine why these mixed results occurred; e.g. due to differences in the samples. Also, future research should investigate why the relationship between personality traits and performance is weaker in introductory accounting courses than in later accounting courses. If it is not performance, then what personality traits determine the self-selection of those proceeding to the later accounting courses and becoming accounting majors and accountants?

Research is also needed on the personality traits, types and temperaments of accounting faculty. Investigation of whether there is a relationship between accounting faculty personality and their effectiveness with certain teaching methods is warranted. Other potential areas include the relationship between accounting faculty personality traits and learning styles of students, and the relationship between accounting faculty personality traits and the accounting courses they teach.

Only two studies have been done in the area of how accountants perform their professional tasks (Chenhall & Morris, 1991; Vassen et al., 1993). These studies looked only at auditors, cognitive processes, and the Sensing-Intuition and Thinking-Feeling scales as non-interactive variables. Research of other professional areas in accounting (tax, consulting, and managerial) is of interest. A strength of personality-based research compared to alternative types of psychology-based research is that aspects of the mind besides cognition and information-processing can be examined; i.e. the mental functions as separate variables. Thus, research using two-way, three-way and four-way preference interactions needs to be done in the various aspects of the profession.

Team or group dynamics is one of the more promising areas of personality-based research (Hammer & Huszczo, 1996). According to the AICPA's Vision (1999), the ability to work in teams, in particular "within diverse, cross-functional teams" is a necessary accounting skill. Team formation is one of the major areas where personality-based instruments are used in the workplace. Research as to the effectiveness of these instruments in this area is needed. Recall that personality-based instruments can examine both cognitive and non-cognitive aspects of teams. Cognitively, teams may be investigated as problem-solving and decision-making units by looking at the distribution of mental functions among team members. Do certain mixes lead to greater efficiency or effectiveness in solving problems? Personality-based instruments are also capable of investigating numerous non-cognitive dimensions of teams. For example, because communication in the team affects overall team performance, are team members with certain combinations of traits better at communicating than others?

Research on team formation, as opposed to performance, using personality-based instruments is needed to investigate the effort management should put into deliberately constructing teams as opposed to allowing teams to self-select. If deliberately forming teams is necessary, then research should focus on which traits and trait interactions are important. Are some membership arrangements better for certain types of tasks than others? For example, research may examine the traits of effective teams as they vary with tax, auditing, financial, or consulting engagements. Similarly, research may indicate that the traits of the client also have an impact on the accounting team's effectiveness.

Furthermore, parallel research questions on teams in the classroom need to be answered using personality-based instruments. Can the accounting instructor improve the effectiveness of team learning and team projects by selecting membership based on personality traits, temperaments or types? Do teams self-select in an equally advantageous manner? Do the personality traits of effective teams varying with the type of accounting course? For example, are the characteristics and skills of effective financial accounting teams different than those of auditing teams?[11]

In addition to being used for capturing the eight traits and 16 personality types, the KTS can be used to specifically investigate hypotheses involving the four temperaments. As discussed above, temperament theory claims to capture important differences between personalities with the four temperaments based on two-way interactions of personality traits (Keirsey, 1998; Berens, 1996, 1998). Since temperament profiles are different from Jungian type profiles, different predictions concerning behavior are expected. Another advantage of the temperament theory approach is that it often requires a smaller sample size than the 16-personality type approach, depending on the hypothesis under consideration.

Numerous MBTI studies have been conducted that include two-way interactions of personality traits among the various combinations of traits examined. These include research into personality characteristics, psychotherapy, health, education, careers, leadership, management and teams (DiTiberio, 1996; Haley, 1997; Hammer, 1996; Kirby & Barger, 1996; Myers & McCaulley, 1985; Myers et al., 1998; Quenk & Quenk, 1996; Shelton, 1996; Walck, 1996). These are not temperament studies per se since they do not look exclusively at the four two-way combinations corresponding to the temperaments. Nor do they employ temperament descriptions in interpreting the results. They do, nevertheless, indicate that there is a great deal of potential for temperament studies because the temperament two-way combinations are among the interactions studied.

CONCLUSIONS

In this paper, three points have been emphasized about temperament theory and the KTS, especially in comparison to Jungian theory and the MBTI. First, temperament theory is unique. Second, the KTS has tremendous potential as a research instrument, similar to the MBTI. Third, while more validity and reliability testing of the KTS is warranted, much has already been done.

In relation to the uniqueness of temperament theory, temperament theory is different from Jungian/MBTI theory. Temperament theory has a different, pre-Jungian history and views Jungian theory as coming, somewhat indirectly, from

temperament theory. There is significant overlap between the two theories. Both theories and their related instruments use the eight Jungian/MBTI personality traits (Extraversion, Introversion, Sensing, Intuition, Thinking, Feeling, Judging and Perceiving) and the concepts of preference and dominance.

The MBTI and KTS that accompany the two theories are similarly related. They both measure the four bipolar scales and accompanying eight traits. However, the KTS was developed specifically to capture the four two-way interactions of the bipolar scales found in temperament theory (i.e. the four temperaments). This is not true of the MBTI, which was developed to capture the four-way interactions (i.e. the 16 types).

The fact that the two instruments employ the same variables to measure the same traits allows researchers to use results from both streams. Although both instruments use the same four scales, how the two theories define the scales are slightly different. Both theories operationalize the same constructs but with slightly different emphases. For example, the KTS "Sensing-Intuition" and the MBTI "Sensing-Intuition" are operationalizations of the same theoretical construct (Sensing-Intuition). The questions used in the two instruments to get at this trait are different, as are the ultimate applications; i.e. determine four temperaments vs. 16 types. This difference is understandable given the highly inclusive nature of the underlying constructs. Personality theorists are trying to capture a very complex phenomenon (the personality) with only four bipolar scales. This partially explains the less than perfect correlation between the two instruments, generally around 0.75.

In relation to KTS' potential as a research instrument, the KTS is a psychometric instrument used to measure constructs (i.e. Extraversion-Introversion; Sensing-Intuition; Thinking-Feeling; Judging-Perceiving) deriving from personality type theory, like the MBTI. Accordingly, researchers may use these measurements in several different ways:

(1) To determine the four two-way interactions known as the temperaments. This is what the KTS was specifically developed for, yet little has been done in this area of accounting.
(2) To determine the four bipolar scales and accompanying eight traits, and use these separately. For example, the Sensing-Intuition and Thinking-Feeling scales have been used separately to operationalize "cognitive style" (e.g. in accounting, Chenhall and Morris 1991). Another example of a study using traits in the accounting profession is the use of the Extraversion-Introversion to predict CPA supervisory opportunities (Satava 1997).
(3) To determine the 16 MBTI/Jungian types. This is not the reason for which the KTS was developed, but the reason for the MBTI. Nevertheless, the

KTS can provide this. Because of the overlap between the KTS and MBTI, researchers using the KTS may draw from and expand on the results from the extensive MBTI research stream.

(4) To obtain multiple measures of the same phenomenon using both the KTS and the MBTI. Multiple measures yield different applications, which advances research and integration across models. Also, they assist individuals in determining their "best fit" type. Evidence of this is apparent in the literature as temperament theory is currently being fused with concepts of systems thinking and multiple intelligences.[12]

In summary, the Keirsey Temperament Sorter is an instrument that allows the accounting researcher to measure certain traits of the personality defined in terms of temperament theory and, indirectly, Jungian personality type psychology. These theories do not limit themselves to any one aspect of the personality, such as information processing, but instead look at the whole personality. The broad scope of temperament theory is an advantage because, through the use of the KTS, it enables the accounting researcher to capture more variation in experimental subjects than that of more specialized psychological paradigms and instruments. There is a rich stream of potential accounting research that can be conducted with the KTS, as illustrated by the small number of accounting articles using personality theories and instruments. The personality types, temperaments and traits of accounting students, teachers, and professionals have been investigated in only a few areas. The MBTI research stream is another rich source from which the accounting researcher can draw examples because the KTS and MBTI have many features in common. Accounting researchers have many opportunities open to them using the Keirsey Temperament Sorter and other Jungian-based instruments.

> It is not the purpose of psychological typology to classify human beings into categories – this in itself would be pretty pointless. Its purpose is rather to provide a critical psychology which will make a methodical investigation and presentation of the empirical material possible. First and foremost, it is a critical tool for the research worker, who needs definite points of view and guidelines if he is to reduce the chaotic profusion of individual experiences to any kind of order. (Jung, 1971, pp. 554–555).

NOTES

1. The literature uses the following standard abbreviations: E for Extraversion, I for Introversion, N for Intuition, S for Sensing, T for Thinking, F for Feeling, J for Judging, and P for Perceiving. For ease of reading, we have avoided using the abbreviations except where necessary. Accordingly, the abbreviations are used in tables and with combinations of traits (e.g. ESTJ for Extraversion-Sensing-Thinking-Judging). We have

retained the convention of capitalizing traits and functions to distinguish the specialized terms from their everyday counterparts – thus, Introversion (the trait) vs. introversion (the common term).

2. The terminology may seem confusing initially. *Traits* refer to individual characteristics such as Extraversion, Introversion, Intuition, Sensing, Thinking, Feeling, Judging or Perceiving. The *bipolar scales* (or *dichotomies*) are the four complementary pairs of the traits; e.g. the Intuition-Sensing pair is the Perceiving bipolar scale. *Types* refer to the four-way combinations or interactions of these traits; e.g. ENTP is a type. Types are usually associated with Jungian personality type theory. *Temperaments* are specific two-way combinations of traits and are usually associated with temperament theory.

3. Temperament theory does not claim to explain all – or even a majority – of the aspects of personality but it does claim to make testable predictions about personality. Like any scientific model, it oversimplifies but hopefully not to the point of being useless for experimental purposes.

4. For an in-depth review of this history, see Keirsey (1998).

5. In should be noted that the MBTI is in a similar situation with regard to its foundations. The theory behind the MBTI derives from Jungian psychology, which is primarily a clinical and observation-based psychology, not an experimental one (Spoto 1989). Jung (1971) states that he can give "no a priori reason for selecting" the divisions he used (p. 437). The input-process-output model used in this paper to explain some of the theoretical constructs are from systems theory and are a latter addition. The experimental support for Jungian theory derives largely from the results obtained from employing the MBTI. Similarly, the experimental support for temperament theory should be based on the effectiveness of the KTS.

6. Jung (1971) explicitly developed a model with 8 types. As discussed further in the Jungian Personality Type Theory section, the model was expanded to 16 types. Most current Jungians accept this expanded model as consistent with ideas implicit in Jung's psychology. Thus, these 16 types are generally referred to as the 16 Jungian types.

7. There is some disagreement among the experts on this point. The orthodox view is that very little change in the personality-type occurs over time. Myers et al. (1998: 164) reports evidence that an informal study conducted over a 50-year period found that 54% of the participants changed either none of their four preferences or only one. The probability of this occurring by chance is 6.25% (Myers et al., 1998). Some, however, hold that radical changes in the individual's personality type can occur during middle age (Mid-life Crisis) (see Pascal, 1992). Another interesting point here is Jung's belief that a perfectly formed individual would possess all 6 traits equally and in balance (Jung, 1971). However, he thought it extremely unlikely that this ever occurs, especially since the personality-structure is strongly set during early childhood. Furthermore, an individual with this ideal personality would probably not have a "personality" or "self" in the traditional sense. The homogeneity of traits would result in an individual without the kinds of differentiation that are used to identify different types of persons.

8. This is a good illustration of Jung's de-emphasis on the unconscious in contrast to Freud.

9. Since the MBTI and KTS are basically measuring the same eight traits or variables (Extraversion, Introversion, Intuition, Sensing, Thinking, Feeling, Judging and Perceiving), the two-way combinations derived from the MBTI should also indicate the corresponding temperaments. There are some slight differences in how the KTS and MBTI define and measure the eight traits, as to be expected since the two instruments

use different sets of questions. Studies between the two instruments in regard to the eight trait variables indicate correlations in the range of 70% (Berens, 1996; Keirsey website, 2002). Thus, at the level of the individual eight variables, the two instruments may be used interchangeably, with some precautions.

10. For the website version, the profile analysis of the four temperaments is free, but the 16 type version using four letters similar to the MBTI is available only if purchased.

11. For examples of personality type-based research in non-accounting areas, see Gardner & Martinko, 1996; Hammer, 1996; and Myers et al., 1998.

12. There are other reasons to consider using the KTS instead of the MBTI. These are primarily logistical and are therefore of secondary importance.

(a) The KTS is shorter. It requires less time, which is often an important consideration when trying to use a psychometric instrument in conjunction with another experimental task.
(b) The online version of the KTS is immediately available. It can be administered in any lab with an Internet connection. The scoring is also done online, relieving the researcher of this task, which may introduce errors. The MBTI must be purchased, installed, have a license agreement, etc.
(c) The researcher does not need to be "qualified" to administer the KTS. MBTI requires researchers to have a certain degree of training.
(d) While the online KTS (KTS website, 2002) is not completely free, it is less expensive than the online MBTI. The KTS website (2002) provides the temperament results free, but full scoring of the eight separate traits (i.e. the four letter version of 16 types) must be purchased. Paper versions of the KTS are also less expensive than those of the MBTI. Furthermore, *Please Understand Me II* (1998) includes a paper copy of the KTS without restricting its use. If the two-way interactions of the four temperaments are all that an individual researcher or educator is interested in (rather that one of the 16 four letter types), then the KTS website (2002) provides this service at no charge.

REFERENCES

Baron, J. (1994). *Thinking and Deciding* (2nd ed.) New York, NY: Cambridge University Press.

Berens, L. (1996). Type and temperament. *Bulletin of Psychological Type, 19*, 8–9.

Berens, L. (1998). *Understanding Yourself and Others: An Introduction to Temperament.* Huntington Beach, CA: Telos Publications.

Berens, L. (1985). A Comparison of Jungian Function Theory and Keirseyan Temperament Theory in the Use of the Myers-Briggs Type Indicator. Unpublished doctoral dissertation, United States International University.

Chenhall, R., & Morris, D. (1991). The effect of cognitive style and sponsorship bias on the treatment of opportunity costs in resource allocation decisions. *Accounting, Organizations and Society, 16*, 27–46.

DiTiberio, J. K. (1996). Education, learning styles, and cognitive styles. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 123–166). Palo Alto, CA: Consulting Psychologists Press.

Feist, J., & Feist, G. J. (1998). *Theories of Personality* (4th ed.). Boston, MA: McGraw-Hill.

Flanagan, O. (1991). *The Science of the Mind* (2nd ed.). Cambridge, MA: MIT Press.

Frisbie, G. R. (1988). Cognitive styles: An alternative to Keirsey's temperaments. *Journal of Psychological Type, 16*, 13–21.

Gardner, W. L., & Martinko, M. J. (1996). Using the Myers-Briggs Type Indicator to study managers: A literature review and research agenda. *Journal of Management, 22*, 45–83.

Gul, F. A., & Fong, S. C. C. (1993). Predicting success for introductory accounting students: Some further Hong Kong evidence. *Accounting Education, 2*, 33–42.

Haley, U. C. V. (1997). The MBTI and decision-making styles: Identifying and managing cognitive traits in strategic decision-making. In: C. Fitzgerald & L. K. Kirby (Eds), *Developing Leaders: Research and Applications in Psychological Type and Leadership Development* (pp. 187–223). Palo Alto, CA: Davies-Black.

Hall, C. S., & Nordby, V. J. (1973). *A Primer of Jungian Psychology*. New York, NY: New American Library, Inc.

Hammer, A. L. (1996). Career management and counseling. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 31–53). Palo Alto, CA: Consulting Psychologists Press.

Hammer, A. L., & Huszczo, G. E. (1996). Teams. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 81–103). Palo Alto, CA: Consulting Psychologists Press.

Harvey, R. J. (1996). Reliability and validity. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 5–29). Palo Alto, CA: Consulting Psychologists Press.

Hergenhahn, B. R., & Olson, M. H. (1999). *An Introduction to Theories of Personality* (5th ed.). Upper Saddle River, NJ: Prentice Hall, Inc.

Hobby, S. A., Miller, D. I., Stone, J. A., & Carskadon, T. G. (1987). An empirical test of differing theoretical positions of Myers and Keirsey concerning type similarity. *Journal of Psychological Type, 13*, 56–60.

Hozik, J., & Wright, J. W., Jr. (1996). A cross-cultural investigation of personality traits among Arab and American business students. *Social Behavior and Personality, 24*, 221–230.

Jung, C. G. (1971[1921]). *Psychological Types*. Princeton, NJ: Princeton University Press.

Kahneman, D., Slovic, P., & Tversky, A. (Eds) (1982). *Judgment under Uncertainty: Heuristics and Biases*. New York, NY: Cambridge University Press.

Keirsey, D. (1998). *Please Understand Me II: Temperament, Character, Intelligence*. Del Mar, CA: Prometheus Nemesis Book Company.

Keirsey, D. (1991). *Portraits of Temperament*. Del Mar, CA: Prometheus Nemesis Book Company.

Keirsey website (2002). The temperaments. URL: http://www.keirsey.com

Kelly, K. R., & Jugovic, H. (2001). Concurrent validity of the online version of the Keirsey Temperament Sorter II. *Journal of Career Assessment, 9* (1, Winter), 49–59.

Kirby, L. K., & Barger, N. (1996). Multicultural applications. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 167–196). Palo Alto, CA: Consulting Psychologists Press.

Laribee, S. (1994). The psychological types of college accounting students. *Journal of Psychological Type, 28*, 37–42.

McCarley, N. G., & Carskadon, T. G. (1986). The perceived accuracy of elements of the 16 type descriptions of Myers and Keirsey among men and women: Which elements are most accurate, should the type descriptions be different for men and women, and do the type descriptions stereotype sensing types? *Journal of Psychological Type, 11*, 2–29.

Meier, C. A., & Wozny, N. A. (1978). An empirical study of Jungian typology. *Journal of Analytical Psychology, 23*, 3–15.

Myers, I. B., & McCaulley, M. (1985). *Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.

Myers, I. B., & McCaulley, M. Quenk, N. L., & Hammer, A. L. (1998). *MBTI Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Myers, I. B., & Myers, P. B. (1995). *Gifts Differing: Understanding Personality Type*. Palo Alto, CA: Davies-Black Publishing.

Nourayi, M. M., & Cherry, A. C. (1993). Accounting students' performance and personality types. *Journal of Education for Business*, (November/December), 111–115.

Nunnally, J. M., & Bernstein, I. (1994). *Psychometric Theory* (3rd ed.). New York, NY: McGraw-Hill Book Company.

Oswick, C., & Barber, P. (1998). Personality type and performance in an introductory level accounting course: A research note. *Accounting Education, 7*, 249–254.

Pascal, E. (1992). *Jung to Live By*. New York, NY: Warner Books, Inc.

Quenk, N. L., & Quenk, A. T. (1996). Counseling and psychotherapy. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 105–122). Palo Alto, CA: Consulting Psychologists Press.

Quinn, M. T., Lewis, R. J., & Fisher, K. L. (1992). A cross-correlation of the Myers-Briggs and Keirsey instruments. *Journal of College Student Development, 33*, 279–280.

Roback, A. A. (1973). *The Psychology of Character*. New York, NY: Arno Press.

Rosenak, C. M., & Shontz, F. C. (1988). Jungian Q-Sorts: Demonstrating construct validity for psychological type and the MBTI. *Journal of Psychological Type, 15*, 33–45.

Ruhl, D. L., & Rodgers, R. F. (1992). The perceived accuracy of the 16 type descriptions of Myers and Keirsey: A replication of McCarley and Carskadon. *Journal of Psychological Type, 23*, 22–26.

Satava, D. (1997). Extroverts or introverts: Who supervises the most CPA staff members? *Journal of Psychological Type, 43*, 40–43.

Shelton, J. (1996). Health, stress, and coping. In: A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 197–215). Palo Alto, CA: Consulting Psychologists Press.

Sipser, M. (1997). *Introduction to the Theory of Computation*. Boston, MA: PWS Publishing Company.

Spoto, A. (1989). *Jung's Typology in Perspective*. Boston, MA: SIGO Press.

Swanger, N. A. (1998). Quick service chain restaurant managers: Temperament and profitability. Unpublished doctoral dissertation, University of Idaho.

Tucker, I. F., & Gillespie, B. V. (1993). Correlations among three measures of personality types. *Perceptual and Motor Skills, 77*, 650.

Tzeng, O. C. S., Ware, R., & Chen, J. (1989). Measurement and utility of continuous unipolar ratings for the Myers-Briggs Type Indicator. *Journal of Personality Assessment, 53*, 727–738.

Vassen, E., Baker, C., & Hayes, R. (1993). Cognitive styles of experienced auditors in the Netherlands. *British Accounting Review, 25*, 367–382.

Walck, C. L. (1996). Management and leadership. In A. L. Hammer (Ed.), *MBTI Applications: A Decade of Research on the Myers-Briggs Type Indicator* (pp. 55–79). Palo Alto, CA: Consulting Psychologists Press.

Ware, R., & Yokomoto, C. (1985). Perceived accuracy of Myers-Briggs Type Indicator descriptions using Keirsey profiles. *Journal of Psychological Type, 10*, 27–31.

Waskel, S. A. (1995). Temperament types: Midlife death concerns, demographics and intensity of crisis. *The Journal of Psychology, 129*, 221–233.

Wheeler, P. R. (2001). The Myers-Briggs Type Indicator and applications to accounting education and research. *Issues in Accounting Education, 16*, 125–150.

Wolk, C., & Nikolai, L. A. (1997). Personality types of accounting students and faculty: Comparisons and implications. *Journal of Accounting Education, 15*, 1–17.