Patrick Griffin · Barry McGaw
Esther Care *Editors*

Assessment and Teaching of 21st Century Skills

Springer

Assessment and Teaching of 21st Century Skills

Patrick Griffin • Barry McGaw • Esther Care
Editors

Assessment and Teaching of 21st Century Skills

Springer

Editors

Patrick Griffin
Melbourne Graduate School of Education
University of Melbourne
Queensberry Street 234
3010 Parkville, Victoria
Australia
p.griffin@unimelb.edu.au

Barry McGaw
Melbourne Graduate School of Education
University of Melbourne
Queensberry Street 234
3010 Parkville, Victoria
Australia

Esther Care
Melbourne Graduate School of Education
University of Melbourne
Queensberry Street 234
3010 Parkville, Victoria
Australia

ISBN 978-94-007-2323-8 e-ISBN 978-94-007-2324-5
DOI 10.1007/978-94-007-2324-5
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2011939474

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

Ubiquitous technology has changed the way people work, live, and play. In contemporary society, people use communication and information technology (ICT) to search for information, make purchases, apply for jobs, share opinions, and stay in touch with friends and relatives. In business, people use technology to work in teams, to create new ideas, products, and services and share these with colleagues, customers, or a larger audience. At the same time, contemporary society faces myriad problems that must be addressed: persistent poverty, HIV/AIDS, food security, energy shortage, global climate change, and environmental degradation. In this context, it is crucial to respond flexibly to complex problems, to communicate effectively, to manage information dynamically, to work and create solutions in teams, to use technology effectively, and to produce new knowledge, continuously. All of these are skills needed in the twenty-first century.

Technology has made profound changes in twenty-first century business and everyday life, but most educational systems operate much as they did at the beginning of the twentieth century. While contemporary business and social practices engage people in collaborative efforts to solve complex problems and create and share new ideas, traditional instructional and assessment practices require students to work individually as they recall facts or perform simple procedures in response to pre-formulated problems within the narrow boundaries of school subjects, and often they do so without the aid of books, computers, social networks, or other resources. School work is shared with and judged by only the teacher and there is little feedback to the student or opportunity for revision. Significant reform is needed in education worldwide: What is learned, how it is learned and taught, and how schools are organized. But reform is *particularly* needed in education assessment and its direct impact on teaching – how it is that education and society, more generally, can advance and measure the competencies, skills, and experiences needed by productive, creative workers and citizens.

Assessments serve an important function when they motivate students to learn, help teachers to refine their practice and develop their skills, and help education systems improve. Assessments can also be used to certify student accomplishments, evaluate the output of educational programs, measure the progress of educational systems, and make comparisons across systems. Most often, this is accomplished with national assessments. But international assessment programs, such as the Programme for International Student Assessment (PISA) and Trends in Mathematics and Science Study (TIMSS), allow countries around the world to compare the performance of their students to other countries and reflect on and improve their educational systems.

But assessment only works if it is measuring the right things. Traditional assessment methods typically fail to measure the high-level skills, knowledge, attitudes, and characteristics of self-directed and collaborative learning that are increasingly important for our global economy and fast-changing world. These skills are difficult to characterize and measure but critically important, more than ever. Traditional assessments are typically delivered via paper and pencil and are designed to be administered quickly and scored easily. In this way, they are tuned around what is easy to measure, rather than what is important to measure. All measure individual results rather than team results. This is no longer acceptable in an economy and society where we need to develop the full potential of all our students.

Insufficient as these assessments are, relative to the needs of our contemporary society and economy, they are one of the most powerful determinants of practice in the classroom, made more so by the use of assessment for high-stakes accountability, where teachers can be fired and schools closed for poor performance. Yet the often-unintended effect of the use of these assessments is to reinforce traditional practices and reduce innovation in schools. Teachers focus on didactic instruction and drill and practice that prepare students for assessments that emphasize the recall of facts and the use of simple procedures. And many previous, well-meaning and well-resourced attempts to reform education have stumbled because they were not able to demonstrate improvement on standardized tests designed for last century's education or because teachers declined to implement them, believing that their students would do poorly on these assessments.

Assessment reform, itself, is a major challenge that requires the efforts, resources, and expertise of not only governments, but industry, academia, as well as non-government institutions. For this reason the three companies – Cisco, Intel, and Microsoft – individually and together, are committed to facilitate research and development to improve education worldwide. They share the belief that high-quality education is important to society and the economy around the world. Each company has an extensive record of support for educational improvement (www.intel.com/education; www.cisco.com/education; www.microsoft.com/education). And together, the companies have worked with UNESCO and the World Economic Forum and other partners to support the development of the UNESCO ICT Competency Standards for Teachers and the Global Education Initiative.

Based on discussions and even direct requests for support from governments and academia, a joint Education Taskforce was set up by the three companies, in the summer of 2008, to review the range of problems, issues, and opportunities in education. The Taskforce chose to target assessment reform as the key factor that will unlock transformation of the educational system across the world. The Taskforce consisted of lead education experts from the three companies (Cisco: Bill Fowler, Andrew Thompson; Intel: Martina Roth, Jon K Price, Lara Tilmanis; Microsoft: Greg Butler, Stephen Coller, Rane Johnson). Dr. Robert Kozma was commissioned to work with the Taskforce in formulating a call to action and initial plans for a joint effort that would support assessment reform. The Taskforce was convinced that assessment reform was a difficult comprehensive challenge that no one segment of the education community or society could resolve on its own, but that requires expertise in measurement, political commitment, academic expertise, technological capability, financial resources, and collaboration with the respective institutions. So the Task Force consulted with policy makers, key academics, and assessment organizations, including experts associated with OECD's Programme for International Student Assessment (PISA) and with the International Association for the Advancement of Educational Achievement. The result was the formulation of the Assessment and Teaching of Twenty-First Century Skills (ATC21S), chaired by Dr. Barry McGaw, University of Melbourne, as Executive Director, and constituted, in its first year, of five Working Groups, that included Twenty-First Century Skills, chaired by Dr. Senta Raizen, WestEd; Methodology, chaired by Dr. Mark Wilson, University of California Berkeley; Technology, chaired Dr. Beno Czapo University of Zeged; Learning Environments, co-chaired by Dr. John Bransford, University of Washington, and Dr. Marlene Scardamalia, University of Toronto; and Policy, chaired by Dr. Linda Darling-Hammond, Stanford University. The Working Groups were charged with analyzing the range of problems that inhibit assessment reform within their specified area and specify potential solutions that can advance assessment reform. Their deliberations included input from over 250 lead researchers across the globe. In addition six pilot countries were identified, with a lead government representative on the Executive Board of the Initiative. An Advisory Board was formed that included the Director of PISA and Chair of IEA, the organization that sponsors TIMSS. The Vice Presidents of Education and Corporate Affairs of Cisco, Intel, and Microsoft expressed their leadership and commitment by chairing the Executive Board of ATC21S (Michael Stephenson, Cisco Corp 2009; Anthony Salcito, Microsoft Corp 2010; Shelly Esque, Intel 2011).

Professor Patrick Griffin of the University of Melbourne was appointed Executive Director of the project at the beginning of 2010 to carry the project forward into its research and development phase. Associate Professor Esther Care also of the University of Melbourne was appointed International Research Coordinator.

This book is the product of phase 1 of the overall ATC21S project. The white papers here have served as the basis for the project's subsequent work in formulating and developing twenty-first century skill assessments. Subsequent phases of the project attempt to add value by catalyzing the international community to identify the opportunities, challenges, issues, and barriers that:

- Are common to all
- Are of the highest priority
- Cannot be addressed by any individual project alone

The intent of the project is *not* to develop an assessment of its own. Rather, the project will provide a structure by which the international community can draw on and share existing knowledge and create effective solutions to address the problems, issues, and barriers associated with the identified skills and foster wide-scale adoption of assessment reforms. All products generated by the project will reside in the public domain.

We offer this collection to you with an invitation to you to join us in advancing this cause. To do so, please visit the project website at http://www.atc21s.org.

Robert B. Kozma
Martina Roth

Contents

List of Figures

List of Tables

Chapter 1
The Changing Role of Education and Schools

Patrick Griffin, Esther Care, and Barry McGaw

Abstract Following a growing awareness that many countries are moving from an industrial-based to information-based economy and that education systems must respond to this change, the Assessment and Teaching of Twenty-First Century Skills Project (ATC21S) was launched at the Learning and Technology World Forum in London in January 2009. The project, sponsored by three of the world's major technology companies, Cisco, Intel and Microsoft, included the founder countries Australia, Finland, Portugal, Singapore and England, with the USA joining the project in 2010. An academic partnership was created with the University of Melbourne. The directorate of the research and development program is situated within the Assessment Research Centre at that university. Two areas were targeted that had not been explored previously for assessment and teaching purposes: Learning Through Digital Networks and Collaborative Problem Solving. The project investigated methods whereby large-scale assessment of these areas could be undertaken in all the countries involved and technology could be used to collect all of the data generated. This in turn was expected to provide data from which developmental learning progressions for students engaged in these twenty-first century skills could be constructed. This project has major implications for teaching and education policies for the future.

Changes in the labour markets in developed economies have changed the skill demands of many jobs. Work environments are technology-rich, problems are frequently ill-defined and people work in teams, often multidisciplinary teams, to deal with them. Major employers bemoan deficiencies in skills in new recruits to their workforces. Cisco, Intel and Microsoft joined forces to sponsor an international, multi-year project to define the skills required in operational terms, to address

P. Griffin (✉) • E. Care • B. McGaw
Melbourne Graduate School of Education, University of Melbourne,
Melbourne, VIC, Australia
e-mail: p.griffin@unimelb.edu.au

P. Griffin et al. (eds.), *Assessment and Teaching of 21st Century Skills*,
DOI 10.1007/978-94-007-2324-5_1, © Springer Science+Business Media B.V. 2012

methodological and technological barriers to their ICT-based assessment and to do this in ways that take account of assessment needs from classroom practice to national and international studies of student achievement. The results of the work will be in the public domain.

Historically, education has responded to and underpinned different forms of power in societies. In the developed Western world, the expansion of education has been strongly associated with the move from agrarian to industrial to information economies. It has fuelled the rise of wealth through industrialisation and led to the 'education of the masses'. Policies of mass education have typically been adopted by countries as they industrialised. Developing nations have sought to replicate these processes and approaches. There is a growing recognition in first world countries, however, that the historical path of advancement may not be the same as the path to future improvement for developing economies.

As technologically advanced nations shift their economies from industrial to information-based, knowledge economies, a number of different systems have emerged across the world. Agrarian economies still exist but in reducing numbers; industrial economies are being replaced but are still essential; information-based economies are increasing, and we are beginning to find combinations of these economic foundations in many developing countries.

The shift from agrarian to industrial production required specific skills both at the level of floor worker and factory supervisor. The shift changed the way people lived and worked, it changed the way people thought, and it changed the kinds of tools they used for work. The new skills and ways of thinking, living and working, once recognised, demanded new forms of education systems to provide them. Similarly, as the products and the technology to develop them become more digitised, another set of management and production skills are needed, focusing on increased digital literacy and numeracy and new ways of thinking. These will increasingly be identified as essential, and pressure on education systems to teach these new skills will intensify. Our lives are already being altered as a result of the shift from industrial to an information-based economy: the ways we work are changing, the ways we think are altering and the tools we use in our employment are almost unrecognisable compared to those that existed 50 years ago. We can anticipate even more of a shift in another 50 years. As global economies move to the trade in information and communications, the demands for teaching new skills will require an educational transformation of a similar dimension to that which accompanied the shift from the agrarian to the industrial era.

With the emergence of the technology-based information age, the role of information in society has changed, and with it, the structure of the workforce. Skilled labour is still important, but a new set of occupations has emerged. Many occupations that depended on the direct use of labour have disappeared. New occupations that depend on information skills have been created. Just as an industrial economy depended on occupations that produced, distributed and consumed products, an information age and a knowledge economy demand occupations that are based on the production, distribution and consumption of information.

Education faces a new challenge: to provide the populace with the information skills needed in an information society. Educational systems must adjust, emphasising information and technological skills, rather than production-based ones.

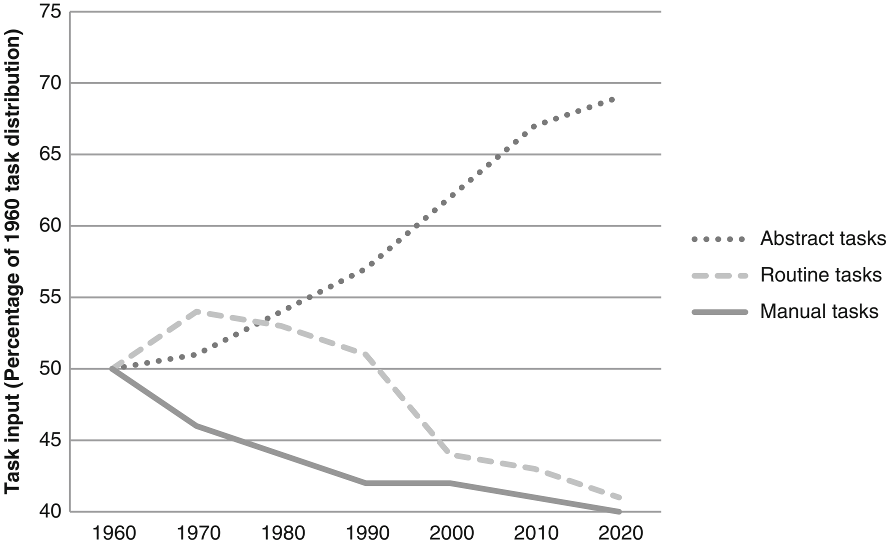


Fig. 1.1 Trends in job tasks (Adapted from Autor et al. 2003)

Those without the skills to act as information producers, distributors and consumers will be disadvantaged, even if their related commodity skills are still in demand. Access to management and advisory roles has become dependent on information skills. The ability to learn, collaborate and solve problems in a digital information environment has become crucial. A study by Autor et al. (2003), shown in Fig. 1.1, illustrates substantial shifts in the structure of the workforce. From 1960 until the present day, there has been an increase in abstract tasks with a corresponding decrease in both routine and manual tasks.

While the nature of education and its role are changing, there is also a need to rethink the way education is measured and monitored. The Organisation for Economic Co-operation and Development (OECD) now examines educational yield in terms of the skills acquired, rather than the number of years of formal education completed. It does this through its Programme for International Student Assessment (PISA). It has done it through its international adult literacy surveys, and is planning to do it through its new Programme for the International Assessment of Adult Competencies (PIAAC) and its planned Assessment of Higher Education Learning Outcomes (AHELO).

This shift illustrates how the meaning of capital has changed in the current age of information and knowledge economies. Power and influence in the industrial age rested on physical capital. This provided a straightforward method of calculating the value of a company, country or social unit: using physical assets. In the information age, human capital is regarded as a means of estimating value. This is due to the perception that capital consists of assets yielding income and other useful outputs over extended periods (Becker 1993). According to this view, expenditure on education and health also represents investment in human capital because they raise earnings, improve health and add to a person's quality of life; investment in education pays dividends because it generates productivity gains. Initially, human capital was measured in terms of years of formal education completed, because there were no comparable metrics of the quality of educational outcomes. Now, the OECD's international measures and those provided by the International Association for the Evaluation of Educational Achievement (IEA) give comparable measures of quality. Within countries, many governments monitor school literacy, numeracy and various other outcomes as measures of human capital. The original measure of human capital (years of formal education completed) has been replaced by an individual's level of literacy and their capacity to access, process, evaluate and use information and to solve problems.

Changing education systems and curriculum to meet the demands of an information and knowledge economy is not enough. Employees also learn and are trained on the job. Regardless of the prerequisite level of education or skills required for any specific employment, employees are typically not fully job ready at the end of their formal education, whether it be secondary or tertiary. Workers often receive additional training to be able to perform their jobs via formal and informal training programmes when they are part of the workforce. Learning increasingly becomes a lifelong process. In a knowledge economy, this is an effect of the shift in the way we learn, the way we think and the way we work. Increased emphasis on technology in the home and the workplace accelerates the need for these new skills.

According to Becker (1993), new technological advances are of little value in countries that have few skilled workers who can use them. Economic growth depends on a synergy between new knowledge and human capital. Hence, countries that have achieved substantial economic growth are those in which large increases in the provision of education and training have been accompanied by advances in knowledge. The information-based role of education in developing twenty-first century skills in an information or knowledge economy has become indisputable.

The ATC21S Project

What are twenty-first century skills? Any skills that are essential for navigating the twenty-first century can be classed as twenty-first century skills. Within the context of the assessment and teaching of twenty-first century skills project (ATC21S), skills so classified must also address the need for, manipulation of and use of information; indeed, they are the primary focus. The ATC21S perspective is that the identified skills do not need to be new. Rather, twenty-first century skills are skills needed and used in the twenty-first century. Some will be familiar and will have been regularly taught and assessed, but essential new skills will emerge.

In the industrial age, categorization of occupations rested on the capacity to develop, distribute and consume products. In the information age a classification of occupations can focus on the production, distribution and consumption of information. This has implications for the outcomes of education. Individuals increasingly need to develop skills for new ways of working, living, learning and thinking. They need new skills to manipulate new information-based work tools.

For example, the need to access and process information in the workplace means that there is an increasing urgency in the need for skills such as analysing the credibility and utility of information, evaluating its appropriateness and intelligently applying it.

These changes in labour markets, especially in developed economies, and where outsourcing of information-based production is preferred, have changed the skill demands of many new jobs. Major employers bemoan the deficiencies in these skills in new recruits to their workforces.

In order to address these issues, three of the world's major technology companies, Cisco, Intel and Microsoft, joined forces to sponsor an international, multi-year project to define the skills required in operational terms, to address methodological and technological barriers to their ICT-based assessment, and to do this in ways that take account of assessment needs from classroom practice to national and international studies of student achievement. They commissioned a paper 'A Call to Action'. Its purpose was to encourage education and government policy makers to respond to the changes technology was having on employment, living and social interaction.

A project, originating from that call to action paper, was designed by a taskforce from the three companies. It was led by Dr Martina Roth of Intel. The taskforce engaged Dr Robert Kozma, formerly of SRI International, to draft the call to action and to develop a detailed proposal. The final design was adopted by the companies and the project was launched at the London Learning and Technology World Forum in January 2009.

The three founding companies negotiated with six national governments to encourage them to join the project as founder countries. These included Australia, Finland, Portugal, Singapore and England, with the USA joining the project in 2010. An academic partnership was created with the University of Melbourne The directorate of the research and development programme is situated within the Assessment Research Centre at that university. Teams were formed in the founder countries. A role for National Project Managers was formulated and national appointments were made in four of the six countries. An executive board was established consisting of the Executive Director, the International Research Coordinator, a Vice President from each of the three companies, and government representatives from the founder countries. An advisory panel was also formed. It consisted of representatives of organisations with global concerns. These included the OECD, the IEA, UNESCO, the World Bank, the Inter-American Development Bank, the National Academy of Sciences and the International Test Commission. Countries that joined the project in its second or third year are represented on the advisory panel.

In the first year of the project, the main products were conceptual documents called white papers. These reviewed previous work and identified issues for research and development. The intended final products were defined to be new assessment strategies and the developmental learning progressions underpinning them that will have been tested and validated in the field in a number of countries. The project's assessment and teaching products will be released into the public domain. The assessment strategies and prototype tasks are to be open-access, open-source, prototype versions.

The White Papers

The first year of the project, 2009, focused on the definitions and parameters of the project. The series of white papers, published in this volume, were commissioned. This stage of the project set out to conceptualise the changes inherent in the shift to an information and knowledge economy, and how this shift would change the way people live and learn, the way they think and work and the tools and procedures used in the workplace. The conceptual structure of the project was organised around these changes in education and skill needs of twenty-first century.

Ways of thinking was conceptualised to include creativity and innovation, critical thinking, problem-solving, learning to learn and the development of metacognition. *Ways of working* was conceptualised to include communication, collaboration and teamwork. *Tools for working* involved information and ICT literacy. *Living in the world* involved changing emphases on local and global citizenship, aspects of life and career development and personal and social responsibility. These were grouped under the acronym KSAVE: knowledge, skills, attitudes, values and ethics. *Ways of learning* and *ways of teaching* are to be considered in the development of the assessment strategies that focus on these skills.

The three companies provided the major component of the project's budget. Founder and associate countries also made a contribution. Five working groups were formed to address the following:

- Identification and definition of twenty-first century skills
- The appropriate methodology of assessment
- The influence of technology on education
- The changes in classroom practice
- The issues of scale and policy development

In addition to the working group leaders, a growing list of researchers became engaged in the work of the project. More than 60 researchers participated in an initial planning conference in San Diego in April 2009. Many others, unable to attend the conference, signalled their interest by engaging with the post-conference work. The OECD and the IEA also became engaged in the work. UNESCO, World Bank and Inter-American Development Bank staff also joined the advisory panel and continue to explore ways in which they might engage with the project. A number of other organisations had the opportunity to join the Advisory Panel. They have done this by proposing to work on particular issues relevant to the project on which they have expertise and for which they might provide funding.

Assessment Development

The ATC21S project is now a multi-year, multinational, public private partnership project that aims to change assessment practices towards a more digital approach using current technology. The project explores changing forms of assessment to match the conceptualisation of twenty-first-century skills. It introduces a methodology for large scale innovative and technology-rich approaches to assessment. As such, it requires a specific project structure, governance, expert panels, field workers and has set out to elaborate on two broad classes of skills that have become the focus of this social, educational and economic change. A new framework for an emerging methodology of assessment development in a technology-rich context has been explored, as have the potential of these changes in assessment to influence education futures. Shifts in thinking about assessment have taken centre stage in this project. The two skills chosen for development (collaborative problem-solving and learning in a digital network) have not been explored previously for assessment and teaching purposes. The approach taken in ATC21S introduces approaches to assessment that involve the deliberate use of ambiguity, a lack of information or definition of problems to be resolved, and interaction between the persons being assessed. It encourages teachers to become involved in the assessment activity with students. Assessment tasks have been developed for a target group of students aged between 11 and 15 years. The data collection process is designed to monitor the way students work together and how they complete a reflective exercise in self and peer assessment.

The Skills Assessed

Collaborative problem-solving has been conceptualised as consisting of five broad strands, the capacity of an individual to: recognise the perspective of other persons in a group; participate as a member of the group by contributing their knowledge, experience and expertise in a constructive way; recognise the need for contributions and how to manage them; identify structure and procedure involved in resolving a problem; and, as a member of the collaborative group, build and develop knowledge and understanding. In the process of developing and field-testing collaborative problem scenarios, broad types of scenarios and tasks are being developed and trialled (Fig. 1.2).

Learning through a digital network has been conceptualised as consisting of the following strands: learning as a consumer of information, learning as a producer of information, learning in the development of social capital and learning in the development of intellectual capital. Again, several broad scenarios are being developed that engage up to four students at a time in identifying procedures and collaborative tools that enable them to learn and develop (Fig. 1.3).

For the two skill areas, tasks have been checked with teachers to ensure that they are realistic, that students will be able to work with them, that the tasks can differentiate between high and low levels of ability, and that the skills underpinning the task resolution are teachable. Think-aloud protocols are being used in small-scale studies (cognitive laboratories) with students representing the target population in order to generate bases for automatic coding and scoring of student performances on the tasks. A series of small-scale pilot studies are also being undertaken in a small number of intact classes to determine the technology and administrative

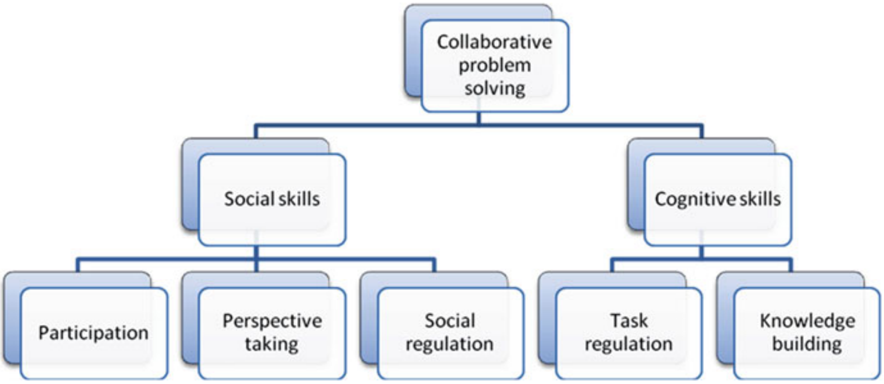


Fig. 1.2 Conceptual framework for collaborative problem-solving (Source: Griffin et al. 2010)

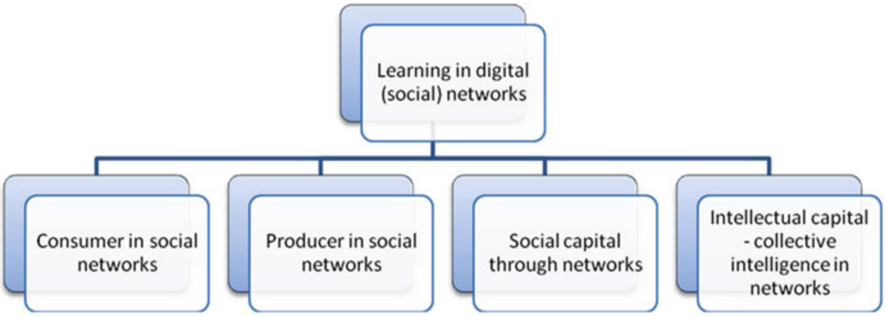

Learning in digital (social) networks
Consumer in social networks
Producer in social networks
Social capital through networks
Intellectual capital - collective intelligence in networks

Fig. 1.3 Conceptual framework for learning in digital networks (Source: Griffin et al. 2010)

requirements for implementation and assessment administration. These represent a rehearsal for the large-scale trials that have been carried out in six countries. Trial data are collected using a matrix-designed sampling approach to identify a cross-national uniform sample of students to maximise calibration accuracy.

These processes are being undertaken with teachers and students in Finland, Singapore, Australia and the United States, as well as associate countries the Netherlands, and Costa Rica, and will be reported in the second volume of this series.

Implications for Pedagogy

One of the more important aspects associated with teaching twenty-first century skills in the ATC21S project is the emergence of a developmental model of learning. It is important to be clear about the difference between deficit and developmental learning approaches, as this difference is central to the mode of twenty-first century teaching. Deficit approaches focus on those things that people cannot do and hence focus on an atomistic ‘fix-it’ perspective. Developmental models build on and scaffold existing knowledge bases of each student and help the student to progress to higher order and deeper levels of learning. A developmental model is also evidence-based and focuses on readiness to learn. It follows a generic thesis of developing the student and points to a way of coping with knowledge explosion in school curricula. Developing twenty-first century skills will require people to work towards higher order thinking and problem-solving. There will be a need for teams of people to work together solving problems who are able to operate at high levels of thinking, reasoning and collaboration. This has implications for teaching as well as for the assessment of these skills. In order to become specialists in developmental learning, teachers need to have skills in using data to make teaching intervention decisions. They will need expertise in developmental assessment, in collaborative approaches to teaching, and a clear understanding of developmental learning models.

In a developmental framework, there is a need to break the ubiquitous link between whole-class teaching and instructional intervention. Teachers will increasingly have to focus on individual developmental and personalised learning for each student. They will also have to work collaboratively rather than in isolation, and base their intervention strategy and resource use decisions on evidence (what students do, say, make or write) rather than inference (what students know, understand, think or feel). When teachers employ a developmental model, their theory of action and psychology of instruction, as well as their thinking, is congruent with theorists who have promoted and given substance to developmental assessment and learning. The teacher’s ability to identify the Vygotskyian (1978) *zone of proximal development* is fundamental to the identification of where a teacher would intervene to improve individual student learning. In order to achieve this with twenty-first century skills, developmental progressions have to be developed and this is a prime goal of the ATC21S project. Teachers need to recognise and use evidence to implement and monitor student progress within a Vygotskyian or developmental approach. Which developmental theory underpins the ATC21S work is negotiable, but choosing a theoretical basis is an important aspect of all forms of teacher education, both pre-service and in-service, if teaching for maximising individual developmental learning in all skill areas is to occur.

When a developmental model of learning is used, the teacher has to reorganise the classroom and manipulate the learning environment to meet the needs of individual students. Manipulation of the learning environment is an important skill. The way in which a teacher links classroom management, intervention strategies and resources used to facilitate learning is always a challenge. The strategies should be guided by a developmental framework of student learning.

Implications for Assessment

There are many stories and studies of the concerns that teachers feel about the emphasis on high-stakes accountability through standardised testing programmes. These programmes help to formulate change in school and higher-level policy and practices, but teachers often feel at a loss with regard to using the data for improvement in classroom teaching and learning. Formative uses of such assessment data generally have not been successful because of the time lag involved in getting data analyses to teachers. This lack of success has led to a generalised shift away from testing and its direct instructional implications. The ATC21S project is developing a different approach to large-scale assessment and reporting to focus as much on direct feedback to teachers and students using individual student data, as it will on informing schools and systems using aggregated data. As such, it may add to the pressure for more direct instruction for pre-service and in-service professional education of teachers in the area of the use of assessment data for instructional purposes.

These changes, however, will require extensive professional education for teachers and for teacher educators. Formal courses in assessment or educational measurement for pre-service teachers are uncommon. The topic ‘assessment’ still conjures up images of multiple choice tests. ‘Tests’ are associated either with standardised measures of literacy and numeracy, or classroom-administered curriculum-based tests of ‘easy to measure’ disciplines. Discussions of standardised measures often evoke normative interpretations, labelling, ranking and deviations. There is a belief that ease of measurement often dictates which subjects are assessed and ‘hard to measure’ subjects are ignored. Assessment and measurement are in turn seen as reducing learning and curriculum to what is easy to measure. In fact, nothing is too hard to measure. As Thurstone (1959) said, ‘If it exists it can be measured, and if it can’t be measured is doesn’t exist’. It all depends on how measurement is defined and how we organise the evidence of more difficult learning concepts. Of course, the core subjects of reading, mathematics and science have been measured for almost a century and the nexus between what is considered important and the skill in measuring them is a solid one. When governments and education systems believe that other skills are as important, the resources and psychometric skill allocation will address student performance in these areas. ATC21S is adding to the list of learning outcomes that have to be considered for their importance. A lot of work is still to be done to convince governments and educators that these new skills deserve large-scale assessment resources and teacher professional development.

Educational measurement demands technical skills. Its specialists are generally engaged in large-scale testing programmes at system, national and international levels. Assessment, on the other hand, requires a different but overlapping set of skills and is linked more generally to teaching and intervention. However, measurement must underpin assessment from a conceptual point of view. Too often at the school level, or in teacher education, measurement or technical aspects of assessment are seen as encroaching on discipline areas of curriculum. Measurement and assessment will increasingly have to refocus on a construct of interest in a developmental framework. Wilson et al. (2011) emphasised this point. It is also argued that assessment *is* a part of curriculum, but it also needs separate, explicit treatment, and educators must develop the relevant skills base. Teachers need the data to make decisions about appropriate intervention, and they need the skills to interpret the implications of data if they are to assist students to develop expertise in twenty-first century skills.

In order to do this, they will need to identify where on a learning progression a student can be located, and in turn there is a need for the ATC21S project to undertake the research in order to define these learning progressions (Wilson et al. 2011). Teachers will have to be convinced of the importance of assessing the skills and developing students along the learning progressions which the ATC21S project initiates.

Policy Implications of Assessment

The process of targeting teaching and focusing on where and how to intervene in developing skills means that there is a need to match strategy with resources and class organisation. There is then a need to coordinate all of this and to implement and evaluate effectiveness. As the effects are identified, issues such as scale and policy need to be reviewed. This can be seen as a policy decision process at the class, school and system levels. At each of the five steps depicted in Fig. 1.4, decisions involve an understanding of the role of time, personnel, materials and space allocation.

Three loops and five steps can be seen in Fig. 1.4. The first loop links measurement directly to intervention. The second loop links resources to policy. The third loop links measurement to policy. The five steps are assessment, generalisation, intervention, resource allocation and policy development. When step two is omitted in the first loop, teachers tend to use an assessment to identify discrete points for teaching. When a test is used without step two, it inevitably leads to teaching what the students cannot do – the deficit model. When the second step (generalisation) is

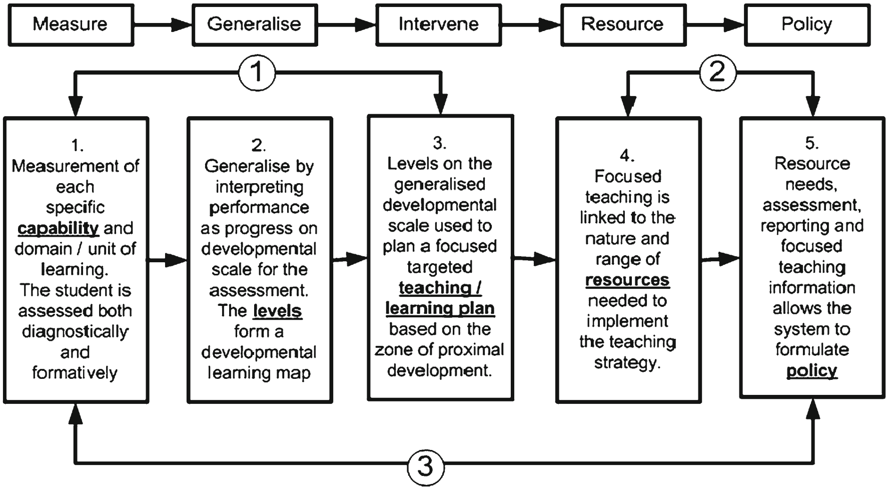

Measure
Generalise
Intervene
Resource
Policy
1
2
1. Measurement of each specific capability and domain / unit of learning. The student is assessed both diagnostically and formatively
2. Generalise by interpreting performance as progress on developmental scale for the assessment. The levels form a developmental learning map
3. Levels on the generalised developmental scale used to plan a focused targeted teaching / learning plan based on the zone of proximal development.
4. Focused teaching is linked to the nature and range of resources needed to implement the teaching strategy.
5. Resource needs, assessment, reporting and focused teaching information allows the system to formulate policy
3

Fig. 1.4 From assessment to policy

included, intervention can be directly linked to a process of teaching to a construct in a developmental approach. On the right of the figure, the link between resources and policy is shown. This is a typical approach for education systems and governments. Resources are the focus of policy formation. The third loop links measurement to policy. The common link is the progression through the five steps which connect learning and policy formation. Progression is achieved by assessing learning in a developmental framework, identifying the generalised level of development, linking resources to the level and intervention strategy, scaling up and formulating policy.

In applying these formative assessment practices, teachers also develop skills in using assessment data to adapt their practices in order to meet students' learning needs. Numerous studies have shown that this is an effective practice in improving teaching and learning (Black and Wiliam 1998; Pressley 2002; Snow et al. 1998; Taylor et al. 2005; Griffin et al. 2010). Assessment data must be based on skills, not scores, and must have the capacity to reflect readiness to learn, rather than achievements or deficits. This is a goal of the ATC21S project: to link assessment with teaching twenty-first century skills.

ATC21S Project Process

The ATC21S project is a research and development project. It has taken assessment and teaching into new territory. The project explores new ideas and skills, new approaches to assessment, and new ways of assessing skills and linking them to teaching interventions aimed at deepening learning and helping to move students to higher order performances. It was planned to consist of five main phases (Fig. 1.5).

The first phase, conceptualisation, was completed in 2009. The result of this was the KSAVE framework and the five white papers. This phase ended in January 2010. In a meeting in London a small number of broad skill areas were identified for further development. These were the areas of collaborative problem-solving and social learning in a digital context.

The second phase was *hypothesis formation*. A second set of expert teams of researchers was recruited from around the world to formulate hypotheses regarding the observable development of 'collaborative problem-solving' and 'learning in a digital network'. In formulating the hypotheses, the teams focused on a number of questions to guide their work:

1. What is the theoretical framework for the construct(s)?
2. What are the purposes of assessing this skill set?
3. What are the functions of this skill set?
4. Is the skill set teachable?
5. Does the skill set form a developmental (non-monotonic) progression?
6. What are the implications and potential for embedding the skill set in a curriculum area?

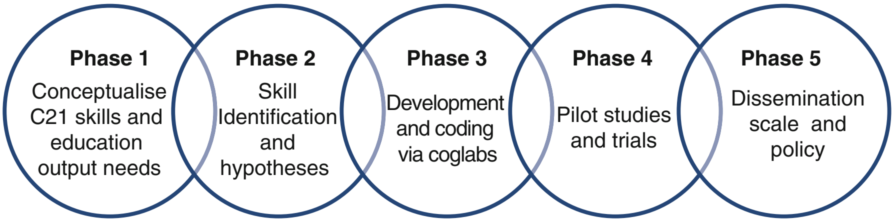

Phase 1
Conceptualise C21 skills and education output needs
Phase 2
Skill Identification and hypotheses
Phase 3
Development and coding via coglabs
Phase 4
Pilot studies and trials
Phase 5
Dissemination scale and policy

Fig. 1.5 The phases of ATC21S project (Source: Griffin et al. 2010)

The third phase of the project involved the development of prototype assessment tasks reflecting the answers to the questions listed above. In the development phase, two steps were used. These were the concept check and the cognitive laboratory.

The purpose of the *concept check* is to check whether teachers considered the early drafts of the tasks relevant and linked to the key learning areas in the curriculum of the participating countries. It was important that this check be undertaken before major task development began. The cognitive laboratory step engaged individual students and teachers in the work of completing the tasks with 'think aloud' and group discussion protocols. The purpose of the *cognitive laboratory* was to identify potential coding categories for automatic scoring and data retrieval.

The fourth phase of the project involves pilot studies and large scale trials of the assessments in order to calibrate them and determine their psychometric properties. The major purpose of the pilot studies is to identify needs such as resources, platforms, administration procedures, time allotment and optimal level of student engagement. The field trials are designed to identify the psychometric properties and calibration of the assessment tasks and to validate the developmental learning progressions. In examining the draft developmental progressions for their utility in a teaching and learning environment, the following questions will be put to teachers about specific students:

1. What is the evidence that could convince a teacher of the location of a student's zone of proximal development?
2. What might be the target for the student and what evidence could convince the team that this is an appropriate target?
3. What teaching strategies or pedagogical approach could be used to enable the student to reach the target?
4. What resource materials would be required?
5. What skills would the teacher have or need to develop in order to move the student forward?

The fifth phase of the project focuses on dissemination. In the final analysis, there is a need to focus on dissemination, implementation, bringing the project outputs and outcomes to scale and helping to formulate policy recommendations. This phase involves the development of materials that will help others to improve on the product and process.

Issues

In addition to the development of the tasks and their conceptual frameworks, there are strategic, technical and perspective issues to be confronted. Large scale assessments of student abilities are relatively common; the focus of the ATC21S project is on skills not yet well understood. This has implications for how teachers understand the constructs which underlie the skills, and how the latter can be enhanced. Without known criteria against which to assess these skills, the project relies on the definitions and the validation of the tasks being developed to justify their importance. As with many innovations, there are tensions between the costs of such a project for its participants and its possible benefits. The capacity of the tasks to lend themselves to a large scale assessment model as well as contribute to the teaching and learning process will be an essential criterion of project success.

Assessment may contribute to driving change – but just one access point, or one driver, is not sufficient. The idea that technology-based large scale assessment will act as ‘a catalyst for a more profound pedagogical change’ (Beller 2011) requires some exploration. There is tension between assessment for change and assessment for identification of current state. Assessment for change informs learning and teaching; assessment for current state informs policy. The nature of the data for these purposes has typically differed. Now we are seeing efforts to use one assessment approach to inform both functions. Whether this is possible without requiring compromises that will diminish the functionality of the assessment for either or both purposes remains to be established. One of the imperatives for ATC21S is to provide both foreground information for use by teachers and background information to harvest for summative system-level analysis.

An assumption of the project is that assessment of twenty-first century skills will lead to a focus on these and contribute to a drive for their inclusion in school curricula. We have seen through national testing practices that assessment can drive teaching in ways that do not necessarily increase student learning. Whether inclusion of assessing ‘skills for living’ might see a similar fate remains to be determined. We know that high-stakes large-scale testing programmes can distort teaching practices, such that teaching to the test replaces teaching to a construct. Teachers have implicitly been encouraged to improve scores but not to improve skills. How do we ensure that systems do not drive such practices? And how do we ensure that teachers understand how to use data from assessment programmes in their teaching? It is essential that teachers are familiarised with the concepts of twenty-first century skills as ‘enabling’ skills in the same way as are literacy and numeracy, if they are to participate in their learning and teaching in a constructive manner. These requirements are at the centre of the ATC21S project’s focus on developmental learning, on assessment tasks which constitute learning tools in their own right, and on the engagement of teachers in the development process.

The expanding list of national and international assessment initiatives that combine aspects of ICT and ‘authentic’ tasks can be seen as a continuation of a traditional approach to assessment, with all its tensions and shortcomings. Although there is a substantial movement toward the use of assessment data for intervention, at the large scale level we have not substantially altered the nature of assessment and appear to think that a traditional approach can fulfil multiple needs. The value of new tools needs to be considered carefully. Think back on your education – what made the most difference? A text you read or a teacher who taught you? The texts and the assessments are tools. We need the workers, and we need workers who know not only how to use the tools but understand the substance with which they are working and the substance with which the learners of today are dealing in the twenty-first century.

These are some of the issues with which ATC21S is engaging, as we move toward large scale assessment with individual scale feedback into the learning loop. In exploring the teaching implications of twenty-first century skills, the project is working closely with teachers, education systems, governments and global organisations represented on the project board and advisory panel in order to link these skills both to new areas of curriculum and to existing discipline-based key learning areas. It is a large and complex undertaking of pioneering work in assessment and teaching of new and previously undefined skills.

References

Autor, D., Levy, F., & Murnane, R. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics, 118*(4), 1279–1333.

Becker, G. (1993). Nobel lecture: The economic way of looking at behavior. *The Journal of Political Economy, 101*(3), 385–409.

Beller, M. (2011). *Technologies in large-scale assessments: New directions, challenges, and opportunities*. International Large Scale Assessment Conference, ETS, Princeton.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139–148.

Griffin, P., Murray, L., Care, E., Thomas, A., & Perri, P. (2010). Developmental assessment: Lifting literacy through professional learning teams. *Assessment in Education: Principles, Policy and Practice, 17*(4), 383–397.

Pressley, M. (2002). Comprehension strategies instruction: A turn-of-the-century status report. In C. C. Block & M. Pressley (Eds.), *Comprehension instruction: Research-based best practices* (pp. 11–27). New York: Guilford.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2005). The CIERA school change fFramework: An evidenced-based approach to professional development and school reading improvement. *Reading Research Quarterly, 40*(1), 40–69.

Thurstone, L. L. (1959). *The measurement of values*. Chicago: The University of Chicago Press.

Vygotsky, L. (1978). *Mind and society: The development of higher psychological processes*. Cambridge: Harvard University Press.

Chapter 2
Defining Twenty-First Century Skills

Marilyn Binkley, Ola Erstad, Joan Herman, Senta Raizen, Martin Ripley, May Miller-Ricci, and Mike Rumble

Abstract As the previous chapter indicates, there has been a significant shift in advanced economies from manufacturing to information and knowledge services. Knowledge itself is growing ever more specialized and expanding exponentially. Information and communication technology is transforming the nature of how work is conducted and the meaning of social relationships. Decentralized decision making, information sharing, teamwork, and innovation are key in today's enterprises. No longer can students look forward to middle class success in the conduct of manual labor or use of routine skills – work that can be accomplished by machines. Rather, whether a technician or a professional person, success lies in being able to communicate, share, and use information to solve complex problems, in being able to adapt and innovate in response to new demands and changing circumstances, in being able to marshal and expand the power of technology to create new knowledge, and in expanding human capacity and productivity.

Research during the last decade has shown how new social practices evolve due to increased use of new digital technologies, especially among young people (Buckingham and Willett 2006). Such practices create reconceptions of

M. Binkley (✉)
University of Luxembourg
e-mail: marilyn.binkley@uni.lu

O. Erstad
University of Oslo

J. Herman
University of California

S. Raizen (retired)

M. Ripley
World Class Arena Limited

M. Miller-Ricci
WestEd, San Francisco, California

M. Rumble
World Class Arena Limited

P. Griffin et al. (eds.), *Assessment and Teaching of 21st Century Skills*,
DOI 10.1007/978-94-007-2324-5_2, © Springer Science+Business Media B.V. 2012

key competencies and skills, not defined from a systems level but from the everyday lives of people in our societies. One example is research done on computer games and online communities (Gee 2007), where problem solving is defined as a key component of such practices. Such experiences of problem solving among young people need to inform us in the way we design assessment tasks and define key competencies. Hence, new standards for what students should be able to do must replace the basic skills and knowledge expectations of the past. To meet this challenge, schools must be transformed in ways that will enable students to acquire the sophisticated thinking, flexible problem solving, and collaboration and communication skills they will need to be successful in work and life. New conceptions of educational standards and assessment, the subject of this chapter, are a key strategy for accomplishing the necessary transformation. Such standards and assessment can both focus attention on necessary capacities and provide data to leverage and evaluate system change. Technology too serves as both a driver and lever for the transformation.

In the sections that follow, we

- synthesize research on the role of standards and assessment in promoting learning,
- describe the nature of assessment systems that can support changes in practice and use these to develop guiding principles for the design of next generation assessments,
- illustrate the use of technology to transform assessment systems and learning, and
- propose a MODEL for assessing twenty-first century skills.

Our intent is to learn from the past as we prepare for new futures in educational standards and assessment. While we provide a list of twenty-first century skills based on our analysis of twelve relevant frameworks drawn from a number of countries, these serve as an example of how to think about assessing twenty-first century skills. We expect that educators, as they consider our model, may need to make adaptations that fit their own contexts as they design assessments appropriate for their schools and students.

We have organized the ten skills we have identified into four groupings:

Ways of Thinking

1. Creativity and innovation
2. Critical thinking, problem solving, decision making
3. Learning to learn, Metacognition

Ways of Working

4. Communication
5. Collaboration (teamwork)

Tools for Working

6. Information literacy
7. ICT literacy

Living in the World

8. Citizenship – local and global
9. Life and career
10. Personal and social responsibility – including cultural awareness and competence

The Role of Standards and Assessment in Promoting Learning

The Importance of Standards That Promote Learning

Worldwide research has established the significant role that curriculum standards and assessment can play in molding new expectations for learning. Although the terminology of standards-led reform may have been initially associated with accountability and improvement initiatives in the USA (e.g., National Center on Education and the Economy 1998; No Child Left Behind Act 2001), the approach has widespread currency in educational systems as divergent as England, Germany, Norway, Singapore, and Australia, to name just a few. The basic ideas followed by these accountability and school improvement systems have rested on three principles:

- Be clear about expectations by establishing standards
- Develop high visibility (sometimes referred to as high stakes) assessments based on the standards
- Use the assessments to communicate what is expected to hold relevant stakeholders accountable and to publish data to inform decisions.

Such standards-based assessments provide empirical evidence for judging performance and can serve a variety of decision-making purposes (accountability, selection, placement, evaluation, diagnosis, or improvement), but the very existence of the assessments and the attention they engender carry important social, motivational, and political consequences.

Researchers around the globe studying such assessments have found fairly uniform effects. This is documented by a number of examples: studies of state accountability assessments in more than a dozen states in the USA, of A- or GCSE or Key Stage Exams in England, and of language and higher education admissions testing programs in countries such as Australia, China, Israel, Japan, New Zealand, and Sri Lanka, and areas such as Central and Eastern Europe (see, for example, Cheng et al. 2004; Herman 2008; Wall 2005). In summary:

- *Assessments signal priorities for curriculum and instruction; high visibility tests serve to focus the content of instruction*. School administrators and teachers pay attention to what is tested, analyze test results, and adapt curriculum and teaching accordingly.
- *Teachers tend to model the pedagogical approach reflected on high visibility tests.* When high visibility assessments are composed of multiple-choice items, teachers tend to rely heavily on multiple-choice worksheets in their classroom

instruction and emphasize lower level cognitive skills. However, when the assessments use extended writing and/or performance assessments, teachers incorporate similar activities in their classroom practice.

- *Curriculum developers, particularly commercial interests, respond* to important tests by modifying existing textbooks and other instructional materials and/or developing and marketing new ones to address test expectations. These products in turn may become primary resources that influence practice and also influence teachers' understandings of test expectations. At the same time research documents effects that can propel productive changes in practice. Thus, it too shows the potential for substantial negative consequences.
- *Schools and teachers tend to focus on what is tested rather than on what the underlying standards or learning goals* are and to ignore what is not tested. Both the broader domain of the tested disciplines and important subjects that are not tested may get short shrift. In the USA, England, and other countries, tests tend to give relatively little attention to complex thinking and problem solving and focus on lower levels of learning, which can lead to similar emphases in classroom practice.
- *Focusing on the test, rather than underlying learning, may encourage a one-time performance orientation and transmission-type teaching.* When doing well on the test, rather than learning, becomes the goal, schools may unwittingly promote a performance orientation in students, which in turn can work against students' engagement and persistence in learning, metacognition, and self-regulation. Especially for high visibility multiple-choice tests, teachers may concentrate on helping students acquire specific content rather than helping students build conceptual understandings and problem-solving capabilities.
- *Instructional/teaching time is diverted to specific test preparation activities*. Schools provide students with practice on the specific types of tasks and formats that are expected on the test through commercial test preparation packages, special classes, and homework. Such activities aim specifically to help students do well on the test, rather than promoting students' learning, and depending on the school and the pressure to improve test scores, can divert weeks or more of instructional time.

These consequences and caveats underscore an important challenge in using assessments to promote twenty-first century skills. The research clearly shows that whatever is measured matters and that educators tend to model and mimic the content and format of high visibility assessments in their curriculum and instruction and use a significant amount of classroom time for special test preparation activities. In some countries, however, testing has become dominated by routine and highly predictable items, which are also often short and highly scaffolded, thus reducing the expectation that students should apply knowledge, skills, and broader capabilities demanded by today's world. For example, analyses of annual state, standards-based tests in the USA show a preponderance of items addressing lower level cognitive demand to the detriment of complex thinking and problem-solving applications (see Webb 1999). Other countries provide more promising examples.

For instance, end of secondary school/university access examinations such as the Baccalaureate, the Matura, Abitur, etc. probe in depth the content and skills that students are expected to acquire and call on students to demonstrate their knowledge and skills in a wide variety of oral and written formats and project-based work. In the Nordic countries, there is a tradition of integrating project work into the curriculum promoting more locally adapted and general standards for assessment. Such examples involve students in important, authentic performances. Even so, the assessment standards for these exams have not yet been fully updated to reflect the demands of an information and innovation age, nor do they take advantage of twenty-first century technology. Just as students need to be literate in new media and be able to harness their power, so too technology can open up new, cost-effective possibilities for the design and use of a new generation of assessments.

Assessment Systems That Promote Learning

The contrast between US-type accountability exams and promising, secondary and university access examinations is also noteworthy in that the latter are embedded in coursework rather than external to it, where they can become an integral part of the teaching and learning process. The exams establish meaningful goals on which course assignments and assessments can be built and are used regularly to assess and respond to student progress. Research shows the powerful effect that ongoing assessment, so-called formative assessment, has on student learning, particularly for low-ability students (Black and Wiliam 1998); OECD 2005).

The use of assessment information is key to the idea: To be considered formative, assessment evidence must be *acted upon* to inform subsequent instruction. Rather than focusing backward on what has been learned, formative assessment helps to chart the learning road forward, by identifying and providing information to fill any gaps between the learners' current status and goals for learning. Moreover, more than solely a source of evidence that informs subsequent teaching and learning, carefully crafted formative assessments can directly support the learning process by incorporating principles of learning and cognition (Herman and Baker 2009; Bennett and Gitomer 2009). For example, by asking students to make public their thinking, formative probes can provide scaffolding that helps students confront their misconceptions, refine and deepen their understandings, and move to more sophisticated levels of expertise (Shepard et al. 2005; Herman and Baker 2005). By asking students for explanations and providing practice over multiple and authentic contexts, assessment tasks can help students to connect new knowledge to their existing structures and build transfer capability (see, for example, Sweller 2003; Holyoak 2005; Ericsson 2002; Gick and Holyoak 1983). By making learning goals explicit and involving students in self-assessment, formative assessment also can promote students as agents in their own learning, increasing student motivation, autonomy, and metacognition, as well as learning (Black et al. 2006; Shepard 2007; Harlen 2006; Gardner 2006). Such characteristics can be similarly incorporated into accountability assessments to increase their learning value.

The Nature of Quality Assessment Systems

Learning-Based Assessment Systems

Assessment design and development must bring together the rich, existing research base on student learning and how it develops with state-of-the-art psychometric theory to produce a new generation of assessments. As a prominent panel in the USA stated:

> Every assessment […] rests on three pillars: a model of how students represent knowledge and develop competence in a subject matter domain; tasks or situations that allow one to observe students' performance; and an interpretation method for drawing inferences from the performance evidence thus obtained (Pellegrino et al. 2001, p. 2).

Adopting this general model, Fig. 2.1 is intended to communicate that quality assessment starts, and ends with clearly specified and meaningful goals for student learning (see also Baker 2007; Forster and Masters 2004; Wilson and Sloane 2000). The assessment task vertex signals that any learning-based assessment must elicit responses that can reveal the quality of student understandings and/or where students are relative to the knowledge and skills that comprise intended learning goals. The interpretation link reinforces the idea that responses from assessment tasks must be specially analyzed and synthesized in ways that reveal and support valid inferences

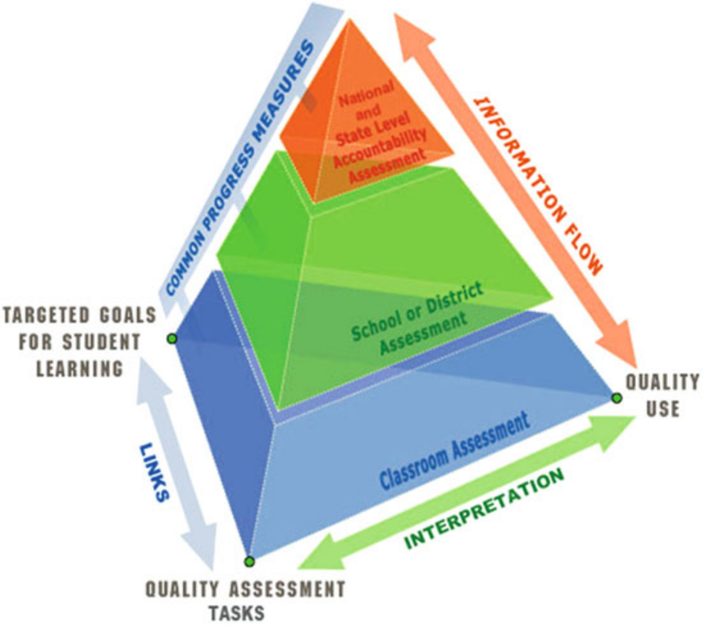

COMMON PROGRESS MEASURES
National and State Level Accountability Assessment
INFORMATION FLOW
School or District Assessment
TARGETED GOALS FOR STUDENT LEARNING
QUALITY USE
Classroom Assessment
LINKS
INTERPRETATION
QUALITY ASSESSMENT TASKS

Fig. 2.1 Integrated assessment system

that connect to intended uses of the assessment. The use vertex highlights that results must be used for student learning relative to initial goals. Assessment quality then resides in the nature of the relationships between and among all three vertices and their connections — in the relationship between learning goals and tasks used to assess their development, in how well the analysis and scoring schemes capture important dimensions of intended understandings and skills, and in how well they support use and are used to improve learning. Inherent here too are the more traditional dimensions of validity, accuracy, and fairness of interpretations of student learning and — particularly for external and higher stakes tests — evidence that interpretations and inferences are justified (see Chap. 3).

As Fig. 2.1 shows, there are multiple levels for which data may be gathered and used for various decision-making purposes, from ongoing data to inform and enrich classroom teaching and learning (see Chap. 5), to periodic data to support policy and practical decision-making at higher levels of the educational system — e.g., school, district, province, state, and national. Importantly, large-scale international, national, and/or state or provincial assessments, for example, may provide policymakers a general barometer for judging and responding to schools' progress in promoting student learning, for allocating resources, and identifying locales that need help, etc. Schools and teachers may use the same data to evaluate their programs, refine their curricula, frame improvement plans, and/or identify individual students who need special attention. But to fuel ongoing decisions to optimize teaching and learning, teachers need a more continuous flow of data. Figure 2.1 implies a system of assessments, grounded in a common, well-specified set of learning goals that is purposively designed to satisfy the decision-making needs of all actors within and across the educational enterprise. Such a system needs to be aligned with the twenty-first century skills that will enable students' future success. Large-scale assessments can serve an important function in communicating and signaling what these skills are, as well as provide important models of how they can be assessed.

Improving the Quality of Assessment Systems

This system perspective also requires a different vantage point for considering assessment quality. Rather than focusing only on a single test, we need to consider the quality of the system for providing valid evidence to support the varied decision-making needs at multiple levels of the educational system. Balanced assessment seems an overriding criterion (Bell et al. 1992). Pellegrino et al. (2001), for example, argued for the development of balanced assessment systems to serve both accountability and policy purposes, as well as those of improving classroom teaching and learning. A balanced system, in their view, incorporates three critical principles: *coherence, comprehensiveness, and continuity*.

- A *coherent* assessment system is built on a well-structured conceptual base — an expected learning progression, which serves as the foundation both for large-scale and classroom assessments. That foundation should be consistent and complementary both across administrative or bureaucratic levels of the education system and across grades.

- A *comprehensive* assessment system uses a range of assessment methods to ensure adequate measurement of intended constructs and measures of different grain size to serve decision-making needs at different levels of the education system. Inherently, a comprehensive assessment system is also useful in providing productive feedback, at appropriate levels of detail, to fuel accountability and improvement decisions at multiple levels.
- *Continuity* captures the principle that assessment at all levels is conceived as part of a continuous stream of evidence that tracks the progress of both individual students and educational programs over time. This can only be possible when there is consistency in the definition of the constructs across time, e.g., from the beginning to the end of the year and across grades.

While inherent in the above formulation, *fairness* is also a fundamental principle for assessment systems. All assessments should be designed to enable the broadest possible population of students to show what they know, without being unfairly hampered by individual characteristics that are irrelevant to what is being assessed. For example, students who are not proficient in the language of the test and test items may well find it difficult to show their mathematics capability; and students from one culture may lack the background knowledge to deal with a reading passage about a context with which they are unfamiliar. Disabled or very-low-ability students may be below the learning threshold on which a test is based. A fair system of assessment offers accommodations for students who may need them and is sensitive to the range of student abilities and developmental levels likely in the assessed population.

Principles for Twenty-First Century Standards and Assessments

While it should be clear that large-scale state, national, regional, or international assessments should be conceived as only part of any system to support student learning, assessments at each level represent a significant opportunity to signal the important learning goals that should be the target of the broader system as well as to provide valuable, actionable data for policy and practice. Moreover, carefully crafted, they can model next generation assessments that, through design and use, can support learning. To do so, our review to this point suggests that twenty-first century standards and assessments should:

- *Be aligned with the development of significant, twenty-first century goals.* Assessments that support learning must explicitly communicate the nature of expected learning. Standards and assessments must fully specify the rich range of twenty-first knowledge and skills students are expected to understand and apply. In addition, the standards and assessments should ideally represent how that knowledge and set of skills is expected to develop from novice to expert performance.

- *Incorporate adaptability and unpredictability.* One hallmark of twenty-first century demands is the need to adapt to evolving circumstances and to make decisions and take action in situations where prior actions may stimulate unpredictable reactions that in turn influence subsequent strategies and options. Dealing with such uncertainty is essential, but represents a new challenge for curriculum and assessment
- *Be largely performance-based.* The crux of twenty-first century skills is the need to integrate, synthesize, and creatively apply content knowledge in novel situations. Consequently, twenty-first century assessments must systematically ask students to apply content knowledge to critical thinking, problem solving, and analytical tasks throughout their education, so that we can help them hone this ability and come to understand that successful learning is as much about the process as it is about facts and figures.
- *Add value for teaching and learning*. The process of responding to assessments can enhance student learning if assessment tasks are crafted to incorporate principles of learning and cognition. For example, assessment tasks can incorporate transfer and authentic applications and can provide opportunities for students to organize and deepen their understanding through explanation and use of multiple representations.
- *Make students' thinking visible*. The assessments should provide a window into students' understandings and the conceptual strategies a student uses to solve a problem. Further, by making students' thinking visible, assessments thus provide a model for quality practice.
- *Be fair.* Fair assessments enable all students to show what they know and provide accommodations for students who would otherwise have difficulty accessing and responding to test items for reasons other than the target of the assessment.
- *Be technically sound.* Assessment data must provide accurate and reliable information for the decision-making purposes for which they are intended to be used. In the absence of reasonable measurement precision, inferences from results, and decisions based on them may well be faulty. The requirement for precision relative to intended purposes means both that intended uses and users must be clearly specified and evidence of technical quality must be established for each intended purpose. Establishing evidence of quality for innovative approaches to assessing twenty-first century skills may well require new psychometric approaches.
- *Valid for purpose.* To the extent an assessment is intended to serve as an indicator of schools' success in helping students acquire twenty-first century skills, skills and test results must be both instructionally sensitive and generalizable. That is, instructionally sensitive tests are influenced by the quality of instruction. Students who receive high-quality instruction should out-perform those who do not. The alternative is that students' basic ability or general intelligence, which are not under a school's control, are the reason for performance. A generalizable result transfers to other real-life applications.
- *Generate information that can be acted upon and provides productive and usable feedback for all intended users.* Teachers need to be able to understand what the assessment reveals about students' thinking. School administrators, policymakers,

and teachers need to be able to use this assessment information to determine how to create better opportunities for student learning.
- *Provide productive and usable feedback for all intended users.* It seems axiomatic that if stakeholders such as teachers, administrators, students, parents, and the public are expected to use the results of an assessment, they must have access to reports that are accurate, understandable, and usable.
- *Build capacity for educators and students.* Feedback from assessments can help students, teachers, administrators, and other providers to understand the nature of student performance and the learning issues that may be impeding progress. Teachers and students should be able to learn from the process.
- *Be part of a comprehensive and well-aligned system of assessments designed to support the improvement of learning at all levels of the educational hierarchy.*

Using Technology to Transform Assessment and Learning

The following sections of this paper address large-scale assessments. Chapter 5 deals more explicitly with classroom assessments.

Assessment Priorities Enabled by Information and Communication Technology

In this section, we draw attention to three areas where ICT has greatly increased the potential for assessing twenty-first century skills. ICT can be thought of not only as a tool for traditional assessments but also as presenting new possibilities for assessing skills formerly difficult to measure. ICT also develops new skills of importance for the twenty-first century. As much as we need to specify the skills needed, we also need to specify approaches that might measure the extent to which students have acquired them. During the last decade, several initiatives have explored how ICT might be used for assessment purposes in different ways in different subject domains. The discussion below is based on a review of relevant research in this area.

Although assessment in education is a substantial research field, it has only been during the last decade that ICT-based assessment has been growing as a research field (McFarlane 2003). This is partly due to an increase in developments of the ICT infrastructure in schools with expanded access to hardware, software, and broadband internet connections for students and teachers. Existing research has examined both the impact of ICT on traditional assessment methods and how ICT raises new issues of assessment and skills. For example, as part of the Second International Technology in Education Study (Kozma 2003), innovative ICT-supported pedagogical practices were analyzed. In several countries, some of these practices demonstrated a shift toward more use of formative assessment methods when ICT was

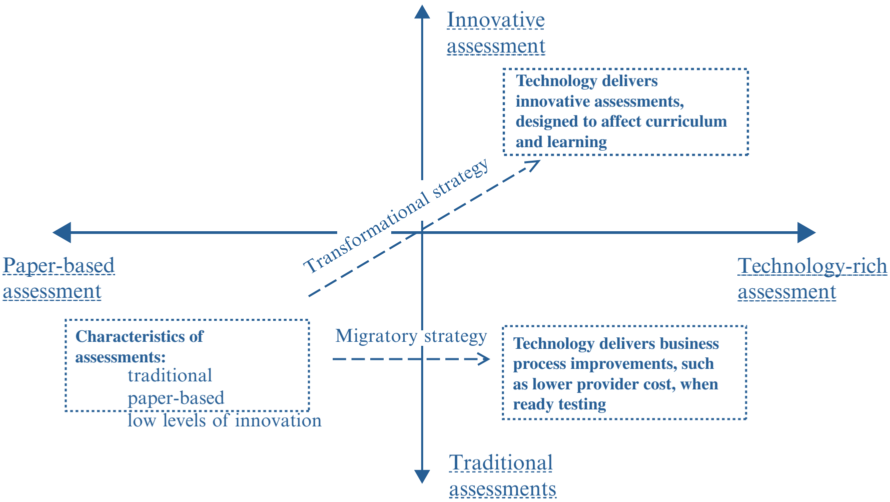

Innovative assessment
Technology delivers innovative assessments, designed to affect curriculum and learning
Transformational strategy
Paper-based assessment
Technology-rich assessment
Characteristics of assessments:
traditional
paper-based
low levels of innovation
Migratory strategy
Technology delivers business process improvements, such as lower provider cost, when ready testing
Traditional assessments

Fig. 2.2 The dimensions of e-assessment innovations

introduced (Voogt and Pelgrum 2003). However, in most practices, often new and old assessment methods coexisted because schools had to relate to national standards and systems over which they had no control, while they were simultaneously developing alternative assessment methods for their own purposes.

The use of the term e-assessment has gained acceptance in recent years. Advocates of e-assessment frequently point to the efficiency benefits and gains that can be realized. These benefits might have to do with the costs of test production, the ability to reuse items extensively or to create power and adaptive tests, or to build system improvements such as test administration systems, which are able to provide tests whenever students want to take them. However, in the report *Effective practice with e-assessment* (Whitelock et al. 2007), the writers conclude that e-assessment is "much more than just an alternative way of doing what we already do." Through evidence and case studies, the report provides examples of e-assessment widening the range of skills and knowledge being assessed, providing unprecedented diagnostic information, and supporting personalization (Ripley 2007). Thus, we argue that e-assessment has the potential of using technology to support educational innovation and the development of twenty-first century skills, such as complex problem solving, communication, team work, creativity and innovation.

Figure 2.2 provides a representation of the contrast between the two drivers: the business efficiency gains versus the educational transformation gains. The lower-left quadrant represents traditional assessments, typically paper-based and similar year-on-year. Most school- and college-based assessments are of this type. Moving from the lower-left to the lower-right quadrant represents a migratory strategy in which paper-based assessments are migrated to a screen-based environment. Delivery is more efficient, but assessments are qualitatively unchanged. In contrast, moving to the upper-right quadrant represents a transformational strategy in which technology is used to support innovative assessment designed to influence (or minimally to reflect) innovation in curriculum design and learning.

The Migratory Strategy with ICT

Conceptions of twenty-first century skills include some familiar skills that have been central in school learning for many years, such as information processing, reasoning, enquiry, critical thinking, and problem solving. The question is: To what extent does ICT enhance or change these skills and their measurement? Indeed, during the last decade most of the research on the use of ICT for assessment has dealt with the improvement of assessment of traditional skills — improvement in the sense that ICT has potential for large-scale delivery of tests and scoring procedures, easily giving the learner accessible feedback on performances. For example, many multiple-choice tests within different subject domains are now online. The focus is then on traditional testing of reasoning skills and information processing among students, on memorization, and on reproduction of facts and information. Using online tests will make this more cost-effective and less time-consuming. However, there are several concerns raised about assessment of traditional skills in an online setting, especially regarding security, cheating, validity, and reliability.

Many countries and states have adopted a "dual" program of both computer-based and paper-and-pencil tests. Raikes and Harding (2003) mention examples of such dual programs in some states in the U.S. where students switch between answering computer-based and paper-and-pencil tests. The authors argue that assessments need to be fair to students regardless of their schools' technological capabilities and the need to avoid sudden discontinuities so that performance can be compared over time. This may require a transitional period during which computer and paper versions of conventional external examinations run in parallel. They sketch some of the issues (costs, equivalence of test forms, security, diversity of school cultures and environments, technical reliability) that must be solved before conventional examinations can be computerized. In a meta-evaluation of initiatives in different states in the US, Bennett (2002) shows that the majority of these states have begun the transition from paper-and-pencil tests to computer-based testing with simple assessment tasks. However, he concludes, "If all we do is put multiple-choice tests on computer, we will not have done enough to align assessment with how technology is coming to be used for classroom instruction" (pp. 14–15).

Recent developments in assessment practices can be seen as a more direct response to the potential of ICT for assessment. An example of such developments is the effort to use computers in standardized national exams in the Netherlands, going beyond simple multiple-choice tests. The domain for the assessment is science, where exams contain 40% physics assignments which have to be solved with computer tools such as modeling, data video, data processing, and automated control technique (Boeijen and Uijlings 2004).

Several studies comparing specific paper-and-pencil testing with computer-based testing have described the latter as highly problematic, especially concerning issues of test validity (Russell et al. 2003). Findings from these studies, however, show little difference in student performance (Poggio et al. 2005), even though there are indications of enough differences in performance at the individual question level to warrant further investigation (Johnson and Green 2004). There are differences in prior computer experience among students, and items from different content areas can be presented and performed on the computer in many different ways, which have different impacts on the validity of test scores (Russell et al. 2003). While some studies provide evidence of score equivalence across the two modes, computerized assessments tend to be more difficult than paper-and-pencil versions of the same test. Pommerich (2004) concludes that the more difficult it is to present a paper-and-pencil test on a computer, the greater the likelihood of mode effects to occur. Previous literature (Russell 1999; Pommerich 2004) seems to indicate that mode differences typically result from the extent to which the presentation of the test and the process of taking the test differ across modes rather than from differences in content. This may imply a need to try to minimize differences between modes. A major concern is whether computer-based testing meets the needs of all students equally and whether some are advantaged while others are disadvantaged by the methodology.

In a recent special issue of the *British Journal of Education Technology* focusing on e-assessment, several studies are presented where students' traditional skills are assessed in different ways (Williams and Wong 2009; Draper 2009; Shephard 2009).

The introduction of ICT has further developed an interest in formative ways of monitoring and assessing student progress. The handling of files and the possibility of using different modes of expression support an increased interest in methods such as project work (Kozma 2003), which can be used for formative assessment. The increased use of digital portfolios in many countries (McFarlane 2003) is an example of how formative assessment is gaining importance. Although the use of portfolio assessments is not new and has been used for some time without ICT (see e.g., special issue in *Assessment in Education*, 1998, on portfolios and records of achievement; Koretz et al. 1998), the use of digital tools seems to have developed this type of assessment further by bringing in some new qualitative dimensions such as possibilities for sending files electronically, hypertexts with links to other documents, and multimodality with written text, animations, simulations, moving images, and so forth. As a tool for formative assessment, and compared to paper-based portfolios, digital portfolios make it easier for teachers to keep track of documents, follow student progress, and comment on student assignments. In addition, digital portfolios are used for summative assessment as documentation of the product students have developed and their progress. This offers greater choice and variety in the reporting and presenting of student learning (Woodward and Nanlohy 2004). This research indicates a strengthening of collaboration (teamwork) and self-regulated learning skills. Related research deals with critical thinking skills, an area of student competency highlighted in curricula in many countries. What is needed in the application of ICT to assessment is to look for new ways of making student attainment visible in a valid and reliable way (Gipps and Stobart 2003; see also Thai school project, critical thinking skills, Rumpagaporn and Darmawan 2007).

In short, in the matter of measuring more traditional skills, development has been directed toward the delivery of large-scale tests on information handling and mapping levels of knowledge at different stages of schooling. Information literacy in this sense has become an important area of competence in itself, and even more so in relation to information sources on the internet. ICT is seen as an important tool in making assessment more efficient as well as more effective in measuring desired skills in traditional ways.

The Transformational Strategy with ICT

Although there are few instances of transformative e-assessment, the projects that do exist provide us with a compelling case for researching and investing in assessments of this type. There are exciting and effective examples of the use of ICT to transform assessment, and, therefore, learning. What is changing in the e-assessment field is usability. Where previously much of the preparatory work had to be done by third party or other technically expert staff, programs are increasingly providing end users with the tools to implement their own e-assessment. New technologies have created an interest in what some describe as "assessing the inaccessible" (Nunes et al. 2003) such as metacognition, creativity, communication, learning to learn, and lifelong learning skills (Anderson 2009; Deakin Crick et al. 2004). Below, we review the research on assessing complex skills that have been difficult to assess or not assessed at all with traditional tests.

The *review of advanced e-assessment techniques* project — commissioned by the Joint Information Systems Committee (JISC) in the UK — began by considering what constituted an advanced technique. "Advanced" refers to techniques that are used in isolated or restricted domains that have successfully applied technology to create an assessment tool. "Advanced" does not necessarily imply "newness." The project collated a long list of over 100 "advanced" e-assessment projects. It was a surprise how few previously unknown advanced e-assessment projects came to light through the trawl for information. The community of experts using e-assessment is small. This continues to have implications for scaling up e-assessment and for stimulating the growth of additional innovative approaches. A brief description of an advanced e-assessment developed in the UK is provided in Fig. 2.3.

One important aspect about the advances in e-assessment is that ICT brings new dimensions to what is being measured. Consider, for example, multimodality, or what Gunter Kress (2003) describes as multimodal literacy. How might different skills like creativity, problem solving, and critical thinking be expressed in different ways using different modes and modalities that ICT provides? The increased uses of visualization and simulation are examples of how ICT has made an impact on measurement of different skills, though so far the research has been inconclusive (Wegerif and Dawes 2004).

Four ICT skills were assessed:

1. Finding things out – obtaining information well matched to purpose by selecting appropriate sources; or, questioning the plausibility and value of information found.
2. Developing ideas and making things happen – using ICT to measure, record, respond to and control events.
3. Exchanging and sharing information – using ICT to share and exchange information, such as web publishing and video conferencing.
4. Reviewing, modifying and evaluating work as it progresses – reflecting critically on own and others' use of ICT.

The design included a simulated environment in which students complete tests; a desktop environment with software and tools for students; new ways of scoring student performances based on the ICT processes students used to solve problems rather than the products, and new ways of enabling access to tests for all students. In one case, an email ostensibly from the editor of a local news website would request students to research local job vacancies and prepare a vacancies page for the website. To complete this task, students would need to run web searches and email virtual companies to request more information about vacancies. The extent and quality of information available would vary, reflecting real-world web information. While completing the task, a student would receive further requests from the editor, perhaps changing deadlines or adding requirements. A student's work would be graded automatically.

The project provided proof-of-concept and identified the following major obstacles and challenges in developing a simulation-based assessment of 21st century skills

- Developing a psychometric approach to measuring and scaling student responses. Since the assessment is designed to collect information about processes used by students, a method is needed to collect data and create summary descriptions/analyses of those processes.
- Aligning schools' technology infrastructure to support wide-scale, high-stakes, computer-based testing.
- Communicating effectively to introduce new approaches to testing to a world of experts, teachers, students, parents and politicians, all of whom have their own mental models and classical approaches for evaluating tests.

Fig. 2.3 Innovative UK assessment of ICT skills of 14-year-olds

Creativity in particular is an area that has been growing in importance as a key twenty-first century thinking skill (Wegerif and Dawes 2004, p. 57). For example, Web 2.0 technology enables users to produce and share content in new ways: User-generated content creation and "remixing" (Lessig 2008) become creative practices that challenge the traditional relationships between teachers and students in providing information and content for learning and the role of the "school book" (Erstad 2008). The use of new digital media in education has been linked to assessment of creative thinking as different from analytic thinking (Ridgway et al. 2004). Digital camera and different software tools make it easier for students to show their work and reflect on it. However, one of the problems with the discussions around creativity has been the often simplified and naïve notions and romantic conceptions of the creative individual (Banaji and Burn 2007), without clear specifications of what this skill area might entail. Thus, it has proved to be difficult to assess students' creativity. In a systematic review of the impact of the use of ICT on students and teachers for the assessment of creative and critical thinking skills, Harlen and Deakin Crick (2003) argue that the neglect of creative and critical thinking in assessment methods is a cause for concern, given the importance of these skills for lifelong learning and in the preparation for life in a rapidly changing society. Their review documents a lack of substantial research on these issues and argues for more strategic research.

A second area of great interest concerns the way digital tools can support collaboration in problem solving, creative practices, and communication. There are many examples of how computer-based learning environments for collaboration can work to stimulate student learning and the process of inquiry (Wasson et al. 2003; Laurillard 2009). Collaborative problem-solving skills are considered necessary for success in today's world of work and school. Online collaborative problem-solving tasks offer new measurement opportunities when information on what individuals and teams are doing is synthesized along the cognitive dimension. Students can send documents and files to each other and, in this way, work on tasks together. This raises issues both for interface design features that can support online measurement and how to evaluate collaborative problem-solving processes in an online context (O'Neil et al. 2003). There are also examples of web-based peer assessment strategies (Lee et al. 2006). Peer assessment has been defined by some as an innovative assessment method, since students themselves are put in the position of evaluators as well as learners (Lin et al. 2001). It has been used with success in different fields such as writing, business, science, engineering, and medicine.

A third area of research with important implications for how ICT challenges assessment concerns higher-order thinking skills. Ridgway and McCusker (2003) show how computers can make a unique contribution to assessment in the sense that they can present new sorts of tasks, whereby dynamic displays show changes in several variables over time. The authors cite examples from the World Class Arena (www.worldclassarena.org) to demonstrate how these tasks and tools support complex problem solving for different age groups. They show how computers can facilitate the creation of micro-worlds for students to explore in order to discover hidden rules or relationships, such as virtual laboratories for doing experiments or games to explore problem-solving strategies. Computers allow students to work with complex data sets of a sort that would be very difficult to work with on paper. Tools like computer-based simulations can, in this way, give a more nuanced understanding of what students know and can do than traditional testing methods (Bennett et al. 2003). Findings such as those reported by Ridgway and McCusker (2003) are positive in the way students relate to computer-based tasks and the increased performances they exhibit. However, the authors also find that students have problems in adjusting their strategies and skills since the assessment results show that they are still tuned into the old test situation with correct answers rather than explanations and reasoning skills.

An interesting new area associated with what has been presented above is the knowledge-building perspective developed by Scardamalia and Bereiter (2006; see also Chap. 5). In developing the technological platform *Knowledge Forum*, Scardamalia and Bereiter have been able to measure students learning processes that have traditionally been difficult to assess. This platform gives the students the possibility of collective reasoning and problem solving building on each other's notes, often as collaboration between schools in different sites and countries. Some key themes in the research on these skills and their online measurement tools are:

- Knowledge advancement as a community rather than individual achievement
- Knowledge advancement as idea improvement rather than as progress toward true and warranted belief

- Knowledge of, in contrast to knowledge about
- Discourse as collaborative problem solving rather than as argumentation
- Constructive use of authoritative information
- Understanding as emergent

Similar points have been made by Mercer and Wegerif and colleagues in the UK (e.g., Mercer and Littleton 2007) in their research on "thinking together" and how we might build language for thinking, what they term as "exploratory talk." Computers and software have been developed for this purpose together with other resources. Wegerif and Dawes (2004, p. 59) have summarized the "thinking together" approach in four points, each of which assumes the crucial importance of teachers:

- The class undertakes explicit teaching and learning of talk skills that promote thinking
- Computers are used both to scaffold children's use of these skills and to bridge them in curriculum areas
- Introductions and closing plenaries are used to stress aims for talk and for thinking as well as to review progress
- Teacher intervention in group work is used to model exploratory talk

The above examples have shown how ICT represents the transformative strategy in developing assessments, especially formative assessment, and how the complexity of these tools can be used to assess skills that are difficult to assess by paper and pencil. As McFarlane (2001) notes, "It seems that use of ICT can impact favorably on a range of attributes considered desirable in an effective learner: problem-solving capability; critical thinking skill; information-handling ability." (p. 230) Such skills can be said to be more relevant to the needs of an information society and the emphasis on lifelong learning than those which traditional tests and paper-based assessments tend to measure.

Arriving at a Model Twenty-First Century Skills Framework and Assessment

In this section, we provide a framework that could be used as a model for developing large-scale assessments of twenty-first century skills. To arrive at this model framework we compared a number of available curriculum and assessment frameworks for twenty-first century skills and skills that have been developed around the world. We analyzed these frameworks to determine not only the extent to which they differ but also the extent to which these frameworks provide descriptions of twenty-first century learning outcomes in measureable form. Based on our analysis, we identified ten important skills that in our opinion typify those necessary for the twenty-first century. For each of the ten skills we have analyzed the extent to which the identified frameworks provide measurable descriptions of the skill, considering the *K*nowledge, *S*kills, and *A*ttitudes, *V*alues and *E*thics aspects of each skill. This framework is referred to as the *KSAVE* framework and is described in more detail below.

Existing Twenty-First Century Skills Frameworks

A number of organizations around the world have independently developed frameworks for twenty-first century skills. For the purposes of our analysis, we considered the frameworks listed in the chart appearing on the next page. To explore the number and range of modern twenty-first century curricula that are currently in place, wider searches were carried out for national education systems that build aspects of the ten KSAVE skills into their national curricula. Searches were made for "national" curricula, references to "twenty-first century learning," and references to "skills" and "competency-based" standards. A relatively small number of nations define a national curriculum in detail, while a larger number have national aims or goals for their education system. A growing number of countries are undertaking significant reviews of their national curricula. A small number are undertaking the task of developing their first national curriculum. "Twenty-first century learning needs" are frequently included within these new and revised curriculum documents. The sources are listed in Table 2.1.

With very few exceptions, references to twenty-first century knowledge, skills, or the individual attitudes and attributes of learners are contained within overarching statements of goals or educational aims. These are generally brief statements but are supported by justifications for change. For example, there are references to: the need to educate for new industry, commerce, technology, and economic structures; the need for new social interaction and communication skills; the need for imagination, creativity, and initiative; the need to learn and continue to learn throughout employment; the need to maintain national and cultural values; and the need to operate in an increasingly international and global environment. Few of the frameworks and curricula of national systems we have examined provide detailed descriptions or clearly elaborated curriculum standards. Similarly, few include descriptions of what the curriculum experienced by learners will actually look like if the broader aims of its framework are to be realized.

All the curricula reviewed maintain a subject structure. It is this structure that forms the basis for curriculum design. The naming and grouping of learning under subject titles may differ slightly between countries, but the general principles of learning a core curriculum (home language, mathematics, and science) are common. In many national curricula, the skills associated with ICT have been raised in status to this core while history, particularly national history, and indigenous culture, often including religion, form a secondary layer. Other subjects may be described individually or combined, for example as the "Arts" or "Humanities." Thus to date the teaching of twenty-first century skills has been embedded in the subjects that make up the school curriculum. It is not clear whether such skills as critical thinking or creativity have features in common in related subjects such as mathematics and science, let alone across the STEM fields and the arts and humanities. For other skills, however, such as information and ICT literacy, the argument has been made more frequently that these are transferrable. These questions of skill generalizability and transferability remain deep research challenges.

Table 2.1 Sources of documents on twenty-first century skills

Country/region	Document(s)
European Union	*Key Competencies for Lifelong Learning – A European Reference Framework, November 2004* Recommendation of the European Parliament and of the Council of 18 December 2006 on key competences for lifelong learning Implementation of "Education and Training 2010" work programme http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:394:0010:0018:en:PDF
OECD	New Millennium Learners Project: Challenging our Views on ICT and Learning www.oecd.org/document/10/0,3343,en_2649_35845581_38358154_1_1_1_1,00.html
USA (partnership for twenty-first century skills)	P21 Framework definitions P21 Framework flyer http://www.p21.org/documents/P21_Framework_Definitions.pdf
Japan	Center for Research on Educational Testing (CRET) www.cret.or.jp/e
Australia	*Melbourne declaration on educational goals for young Australians* www.mceecdya.edu.au/verve/_resources/National_Declaration_on_the_Educational_Goals_for_Young_Australians.pdf
Scotland	A curriculum for excellence – the four capabilities www.ltscotland.org.uk/curriculumforexcellence/index.asp
England	*The learning journey*
England	*Personal learning & thinking skills – the national curriculum for England* http://curriculum.qcda.gov.uk/uploads/PLTS_framework_tcm8-1811.pdf
Northern Ireland	*Assessing the cross curricular skills* http://www.nicurriculum.org.uk/key_stages_1_and_2/assessment/assessing_crosscurricular_skills/index.asp
ISTE	*National educational technology standards for students, second edition, global learning in the digital age* http://www.iste.org/standards.aspx
USA. National Academies, science for the twenty-first century	*Exploring the intersection of science education and the development of* twenty-first *century skills* http://www7.nationalacademies.org/bota/Assessment_of_21st_Century_Skills_Homepage.html
USA, Department of Labor	Competency models: *A review of the literature* *The role of the Employment and Training Administration* (ETA), Michelle R. Ennis

Where the aims and goals of twenty-first century learning are described in the frameworks we examined, they are generally specified as being taught through, within and across the subjects without the detail of how this is to be achieved or what the responsibilities of each subject might be in achieving them. Without this depth of detail, these national statements of twenty-first century aims and goals are unlikely to be reflected in the actual learning experience of students or in the assessments that are administered. Without highly valued assessments of these twenty-first century aims or goals requiring their teaching, it is difficult to see when or how education systems will change significantly for the majority of learners.

The KSAVE Model

To structure the analysis of twenty-first century skills frameworks, an overall conceptual diagram was created. This diagram defines ten skills grouped into four categories:

Ways of Thinking

1. Creativity and innovation
2. Critical thinking, problem solving, decision making
3. Learning to learn, metacognition

Ways of Working

4. Communication
5. Collaboration (teamwork)

Tools for Working

6. Information literacy (includes research on sources, evidence, biases, etc.)
7. ICT literacy

Living in the World

8. Citizenship – local and global
9. Life and career
10. Personal and social responsibility – including cultural awareness and competence

Although there are significant differences in the ways in which these skills are described and clustered from one framework to another, we consider that the above list of ten is sufficiently broad and comprehensive to accommodate all approaches. At an early stage we found that frameworks for twenty-first century skills differ considerably in terms of the nature of their content. Some seek to define student behaviors; for example, an aspect of creativity might include “openness and responsiveness to new ideas.” Other frameworks refer extensively to skills: for example, an aspect of creativity might refer to the ability to “develop innovative and creative ideas.” A third category used in some frameworks refers to specific knowledge: for example, an aspect of creativity might be “knowledge of a wide range of idea creation techniques.” Some frameworks cover two or more of these categories; few comprehensively cover all three. To accommodate and reflect these differences in approach, we have designed three categories within the KSAVE model. Keep in mind that the model does not resolve the issue of subject-embedded knowledge, skills, and attitudes versus their generalizability across domains.

Knowledge

This category includes all references to specific knowledge or understanding requirements for each of the ten skills.

Skills

This category includes the abilities, skills, and processes that curriculum frameworks are designed to develop in students and which are a focus for learning.

Attitudes, Values, and Ethics

This category refers to the behaviors and aptitudes that students exhibit in relation to each of the ten skills.

The method used to complete the analysis of twenty-first century skills frameworks was to populate the KSAVE grid with indexes taken from each framework, retaining original wording as far as was sensible. Decisions were made to refine or amalgamate wording taken from frameworks where the intention appeared similar. Decisions were also made on whether to allocate indexes to knowledge, skills, or attitudes/values/ethics. For some of the indexes, the decision whether to allocate them to the skills category or to the attitudes/values/ethics category appeared to be marginal.

In the following pages, we present each group of skills and discuss some of the thinking behind the grouping. In addition, we provide examples of how the skills might be measured in an effort to open our eyes to what is possible. These example assessments really only scratch the surface of what is needed to measure twenty-first century skills.

Ways of Thinking

Together the three categories of skills under “Ways of thinking” represent a push forward in the conceptualization of thinking. These skills emphasize higher order thinking skills, and subsume more straightforward skills such as recall, and drawing inferences. A major characteristic of these skills is that they require greater focus and reflection.

Creativity and Innovation

Operational definitions of creativity and innovation are provided in Table 2.2. While creativity and innovation can logically be grouped together, they originate in two different traditional schools of thought. Creativity is most often the concern of

Table 2.2 Ways of thinking – creativity and innovation

Knowledge	Skills	Attitudes/values/ethics
Think and work creatively and with others • Know a wide range of idea creation techniques (such as brainstorming) • Be aware of invention, creativity, and innovation from the past within and across national boundaries and cultures • Know the real-world limits to adopting new ideas and how to present them in more acceptable forms • Know how to recognize failures and differentiate between terminal failure and difficulties to overcome *Implement innovations* • Be aware of and understand where and how innovation will impact and the field in which the innovation will occur • Be aware of the historical and cultural barriers to innovation and creativity	*Think creatively* • Create new and worthwhile ideas (both incremental and radical concepts) • Be able to elaborate, refine, analyze, and evaluate one's own ideas in order to improve and maximize creative efforts *Work creatively with others* • Develop, implement, and communicate new ideas to others effectively • Be sensitive to the historical and cultural barriers to innovation and creativity *Implement innovations* • Develop innovative and creative ideas into forms that have impact and can be adopted	*Think creatively* • Be open to new and worthwhile ideas (both incremental and radical) *Work creatively with others* • Be open and responsive to new and diverse perspectives; incorporate group input and feedback into the work • View failure as an opportunity to learn; understand that creativity and innovation is a long-term, cyclical process of small successes and frequent mistakes *Implement innovations* • Show persistence in presenting and promoting new ideas

cognitive psychologists. Innovation, on the other hand, is more closely related to economics where the goal is to improve, advance, and implement new products and ideas. Measuring both can be quite challenging. The tasks require an interactive environment, but they frequently cannot be done in the short period of time allocated to a large-scale assessment, nor are there good benchmarks against which respondent output can be evaluated.

Creativity is often described as a thinking skill or at least as an important aspect of thinking that can and should be fostered (Wegerif and Dawes 2004, p. 57). In a review of the connection between technology, learning, and creativity, Loveless (2007) shows how technology allows children to produce high quality finished products quickly and easily in a range of media that provide opportunities for creativity. Loveless argues that to foster creativity in the classroom, teachers need to create a social atmosphere in which children feel secure enough to play with ideas and to take risks.

Although, as noted above, it has proven to be difficult to assess creativity, the use of new digital media has been linked to assessment of creative thinking as different from analytic thinking (Ridgway et al. 2004). Digital cameras and different software tools make it easier for students to show their work and reflect on it. A number of subjects in the school curriculum ask students to make various kinds of products. (Sefton-Green and Sinker 2000). These might include paintings in art class, creative writing in english, performance in drama, recording in music, videos in media studies, and multimedia "digital creations" in different subjects. There are so far not many examples of how ICT might influence assessment of such student products (Sefton-Green and Sinker 2000).

eSCAPE

The eSCAPE project does not test creativity and innovation, but it does test some aspects of this domain. Specifically it offers a glimpse of how we might test the ability to develop innovative and creative ideas into forms that have impact as well as showing persistence in presenting and promoting new ideas.

For many years, England's school examinations for 16-year-old students have included an optional assessment in Design and Technology. Traditionally, this examination includes a requirement for students to complete a design project of over 100 h duration and to submit a written report on the project. The report is graded.

In 2001, the Qualifications and Curriculum Authority commissioned the Technology Education Research Unit (TERU) at Goldsmiths College in London to undertake to develop a technology-led replacement to this traditional paper-based assessment. The result is an assessment completed in six hours, in a design workshop, with students working in groups of three or four. During the course of the six h, students are given a number of staged assessment instructions and information via a personal, portable device. The handheld device also acts as the tool to capture assessment evidence – via video, camera, voice, sketchpad, and keyboard. During the six hours, each student's design prototype develops, with the handheld device providing a record of progress, interactions, and self-reflections.

At the end of the assessment, the assessment evidence is collated into a short multimedia portfolio. Human raters, who score each student's responses, view this. eSCAPE directors turned to the work of Thurstone (1927) to develop a graded-pairs scoring engine to provide a holistic judgment on the students' work. This engine supports human raters in making a number of paired judgments about students' work. The result is an assessment that exhibits rates of reliability equal to, or slightly in excess of, the levels of reliability achieved on multiple-choice tests.

Critical Thinking, Problem Solving and Decision Making

Operational definitions of critical thinking and problem solving are provided in Table 2.3. Critical thinking and problem solving have become an increasingly important feature of the curriculum in many parts of the world. In the UK there are popular high school qualifications in critical thinking. In the USA, the American Philosophical Association has published the Delphi report on critical thinking (Facione 1990). This report identified six cognitive thinking skills: interpretation, analysis, evaluation, inference, explanation, and self-regulation. This framework was further elaborated

Table 2.3 Ways of thinking – critical thinking, problem solving, and decision making

Knowledge	Skills	Attitudes/values/ethics
Reason effectively, use systematic thinking and evaluate evidence • Understand systems and strategies for tackling unfamiliar problems • Understand the importance of evidence in belief formation. Reevaluate beliefs when presented with conflicting evidence *Solve problems* • Identify gaps in knowledge • Ask significant questions that clarify various points of view and lead to better solutions *Articulation* • Clearly articulate the results of one's inquiry	*Reason effectively* • Use various types of reasoning (inductive, deductive, etc.) as appropriate to the situation *Use systems thinking* • Analyze how parts of a whole interact with each other to produce overall outcomes in complex systems. Examine ideas, identify, and analyze arguments • Synthesize and make connections between information and arguments • Interpret information and draw conclusions based on the best analysis. Categorize, decode, and clarify information • Effectively analyze and evaluate evidence, arguments, claims, and beliefs • Analyze and evaluate major alternative points of view. • Evaluate. Assess claims and arguments • Infer. Query evidence, conjecture alternatives, and draw conclusions • Explain. State results, justify procedures, and present arguments. • Self-regulate, self-examine, and self-correct	*Make reasoned judgments and decisions* • Consider and evaluate major alternative points of view • Reflect critically on learning experiences and processes • Incorporate these reflections into the decision-making process *Solve problems* • Be open to non-familiar, unconventional, and innovative solutions to problems and to ways to solve problems • Ask meaningful questions that clarify various points of view and lead to better solutions *Attitudinal disposition* • Trustful of reason • Inquisitive and concerned to be well informed • Open and fair minded • Flexible and honest • Inquisitiveness and concern to be well informed • Alert to opportunities to use ICT • Trustful of and confident in reason • Open and fair minded, flexible in considering alternative opinions • Honest assessment of one's own biases • Willingness to reconsider or revise one's views where warranted

to include attitudinal- and values- based criteria: Students should be inquisitive, well informed, open-minded, fair, flexible, and honest. Research subsequent to the Delphi Report has shown that being "trustful of reason" (one of the Delphi Report's key findings) plays a vital role in what it means to think critically.

In contrast to creativity and innovation, critical thinking, problem solving, and decision making have been part of large-scale assessments for some time. Critical thinking frequently appears as part of reading, mathematics, and science assessments, with such assessments as the US National Assessment of Educational Progress and the OECD Program for International Student Achievement (PISA).

Problem solving has been a focused area of research for decades, yielding a number of definitions and frameworks. In addition, problem solving has appeared in various forms in a number of large-scale international assessments such as PISA and the Adult Literacy and Lifelong Learning Skills (ALL). These assessments specifically include items that are designed to measure how well students can evaluate evidence, arguments, claims, and warrants; synthesize and make connections between information and arguments; and analyze and evaluate alternative points of view. ALL 2003 focused on problem-solving tasks that were project oriented and most closely resembled analytic reasoning. Problem solving in mathematics and science has been part of the PISA assessment since its inception in 2000. In PISA 2003 a problem-solving scale that included three kinds of problems – decision-making, system analysis and design (and troubleshooting) was developed. For 2012, PISA will move beyond the 2003 scale by including dynamic items that may be linked to the OECD's Program for the International Assessment of Adult Competencies (PIAAC) 2011, where problem solving is in a technology rich environment is measured.

The following examples illustrate the direction of assessments for the twenty-first century. The first, Primum, from the USA, illustrates authentic open-ended tasks that can be machine scored. The second example, World Class Tests, illustrates highly innovative problem solving in mathematics, science, and design and technology that are by design not familiar to the student (much of our current testing is routine and predictable), interesting, motivating, psychologically challenging, and focused on a specific dimension of problem solving, such as optimization or visualization, in a mathematics/science/design context. These tasks offer the hope that it is possible to design lively, 5–10 min long, interactive, and complex problems for students to solve in the context of an on-screen test. The third example, the Virtual Performance Assessment (VPA) project, also from the USA, addresses the feasibility of using immersive technologies to deliver virtual performance assessments that measure science inquiry knowledge and skills, as defined in the U.S. National Science Education Standards (NRC 1996).

Primum

Some advocates of e-assessment point to the potential of computers to support simulation and scenario-based assessment. There are few examples of this category of e-assessment being developed successfully, especially not in high-stakes testing contexts. Primum, which assesses decision making in a very specific context, is an exception. It provides an assessment of trainee medical practitioners' ability to make medical diagnoses when presented with a fictitious patient exhibiting a number of symptoms. This automated assessment has been designed to provide an authentic and reliable assessment at a price that compares favorably with the alternative – human-scored evaluation at patients' bedsides.

World Class Tests

In 2000, England's Department for Education commissioned the development of new computer-based tests of problem solving in the domains of mathematics, science, and design and technology. These tests are intended for worldwide application and were designed to make creative use of computer technology. Also, they are intended to set new benchmarks in the design of assessments of students' thinking and ability to apply a range of techniques to solve novel and unexpected problems. These tests have become known as World Class Tests and have been adapted for children aged 8–14. These tests are now sold commercially under license in East Asia.

The VPA Project

The Virtual Performance Assessment project utilizes innovations in technology and assessment to address the problem of measuring a student's ability to perform scientific inquiry to solve a problem. The project is developing assessments for use in school settings as a standardized component of an accountability program. The goal is to develop three assessments in the context of life science that appear different on the surface, but all measure the same inquiry process skills. Each assessment will take place in a different type of ecosystem, and students will investigate authentic ecological problems as they engage in the inquiry process.

Learning to Learn and Metacognition

Operational definitions of learning to learn and metacognition are provided in Table 2.4. Learning to learn and metacognition have most frequently been measured by think-aloud protocols that have been administered in one-on-one situations. Clearly this methodology is not amenable to large-scale assessments. However, technology might be used to support and assess learning to learn, which includes self-assessment and self-regulated learning. One interesting example of this is the eVIVA project developed at Ultralab in the UK.

eVIVA

The intention of eVIVA was to create a more flexible method of assessment, taking advantage of the possibilities new technologies such as a mobile phone and web-based formative assessment tools offer. By using such tools, project authors Ultralab promoted self- and peer-assessment as well as dialogue between teachers and students.

Table 2.4 Ways of thinking – learning to learn, metacognition

Knowledge	Skills	Attitudes/values/ethics
• Knowledge and understanding of one's preferred learning methods, the strengths and weaknesses of one's skills and qualifications • Knowledge of available education and training opportunities and how different decisions during the course of education and training lead to different careers	• Effective self-management of learning and careers in general. Ability to dedicate time to learning, autonomy, discipline, perseverance, and information management in the learning process • Ability to concentrate for extended as well as short periods of time • Ability to reflect critically on the object and purpose of learning • Ability to communicate as part of the learning process by using appropriate means (intonation, gesture, mimicry, etc.) to support oral communication as well as by understanding and producing various multimedia messages (written or spoken language, sound, music etc.)	• A self-concept that supports a willingness to change and further develop competencies as well as motivation and confidence in one's capability to succeed • Positive appreciation of learning as a life-enriching activity and a sense of initiative to learn • Adaptability and flexibility • Identification of personal biases

In this project, the students had access to the eVIVA website where they could set up an individual profile of system preferences and record an introductory sound file on their mobile phone or landline. After this, students could carry out a simple self-assessment activity by selecting a series of simple "I Can" statements designed to start them thinking about what they are able to do in ICT. The website consisted of a question bank from which the pupils were asked to select four or five questions for their telephone viva or assessment carried out toward the end of their course, but at a time of their choice. Students were guided in their choice by the system and their teacher. They had their own e-portfolio web space in which they were asked to record significant *milestone* moments of learning and to upload supporting files as evidence. Each milestone was then annotated or described by the pupil to explain what they had learned or why they were proud of a particular piece of work. Once milestones had been published, teachers and pupils could use the annotation and the messaging features to engage in dialogue with each other about the learning. Students were encouraged to add comments to their own and each other's work. The annotations could be sent via phone using SMS or voice messages. When ready, students would dial into eVIVA and record their answers to their selected questions. This gave students the opportunity to explain what they had done and reflect further on their work. Their answers were recorded and sent to the website as separate sound files. The teacher made a holistic assessment of the pupil's ICT capabilities based on the milestones, work submitted in the e-portfolio, student reflections or annotations, the recorded eVIVA answers, any written answers attached to the questions, and classroom observations (see Walton 2005).

Cascade

Cascade, which is under development at the University of Luxembourg and the Center for Public Research Henri Tudor, is an innovative item type that is more amenable to large-scale assessments with limited testing time.

The Cascade test items are designed so that respondents answer a set of questions and are then asked to rate how certain they are about the correctness of their response on each item. Then the respondent is given an opportunity to access multimedia information to verify the correctness of the response. At that point, the respondent once again answers the same set of questions and again rates his/her certainty. Scoring is based on the comparison of the first and second set of responses and tracing the information information paths he/she took in acquiring additional information.

Ways of Working

In business, we are witnessing a rapid shift in the way people work. Outsourcing services across national and continental borders are just one example. Another is having team members telecommute while working on the same project. For instance, a small software consulting team has members located in three continents. They work on developing prototypes using teleconferences and email, with the occasional "sprint" sessions where they gather in a single location and work 24 h a day to develop the product. Similarly, in the large-scale international assessments such as PISA, TIMSS (Trends in Mathematics and Science Study), and PIAAC, teams of researchers and developers across continents and at multiple locations work together to develop the assessments. To support these examples of moves toward globalization, communication and collaboration skills must be more finely honed. Communication must be rapid, concise, and cognizant of cultural differences.

Communication

Operational definitions of communication are provided in Table 2.5. Communication has been a mainstay of assessments in the form of reading, writing, graphing, listening and speaking. However, the assessments have not taken into account the full range of possibilities. At the most minimal, PowerPoint presentations are now ubiquitous. These frequently include graphic displays that, in conjunction with language, can more succinctly deliver a message. Video presentations also require the combination of communication forms in ways that have never before been within the realm of most people's capability. To date, newer modes of communication have rarely been represented in large-scale assessments. However, in light of the developments described below, it is essential that we take these changes into account.

Table 2.5 Ways of working – communication

Knowledge	Skills	Attitudes/values/ethics
Competency in language in mother tongue. • Sound knowledge of basic vocabulary, functional grammar and style, functions of language • Awareness of various types of verbal interaction (conversations, interviews, debates, etc.) and the main features of different styles and registers in spoken language • Understanding the main features of written language (formal, informal, scientific, journalistic, colloquial, etc.) *Competency in additional language/s.* • Sound knowledge of basic vocabulary, functional grammar and style, functions of language • Understanding the paralinguistic features of communication (voice-quality features, facial expressions, postural and gesture systems) • Awareness of societal conventions and cultural aspects and the variability of language in different geographical, social, and communication environments	*Competency in language in mother tongue and additional language/s.* • Ability to communicate, in written or oral form, and understand, or make others understand, various messages in a variety of situations and for different purposes • Communication includes the ability to listen to and understand various spoken messages in a variety of communicative situations and to speak concisely and clearly • Ability to read and understand different texts, adopting strategies appropriate to various reading purposes (reading for information, for study, or for pleasure) and to various text types • Ability to write different types of texts for various purposes and monitor the writing process (from drafting to proofreading) • Ability to formulate one's arguments, in speaking or writing, in a convincing manner and take full account of other viewpoints, whether expressed in written or oral form • Skills needed to use aids (such as notes, schemes, maps) to produce, present, or understand complex texts in written or oral form (speeches, conversations, instructions, interviews, debates)	*Competency in language in mother tongue.* • Development of a positive attitude to the mother tongue, recognizing it as a potential source of personal and cultural enrichment • Disposition to approach the opinions and arguments of others with an open mind and engage in constructive and critical dialogue • Confidence when speaking in public • Willingness to strive for aesthetic quality in expression beyond the technical correctness of a word/phrase • Development of a love of literature • Development of a positive attitude to intercultural communication *Competency in additional language/s.* • Sensitivity to cultural differences and resistance to stereotyping

Consider the use of text messaging. The first commercial text message was sent in December of 1992. Today the number of text messages sent and received everyday exceeds the total population of the planet. Facebook, which started as a communication vehicle for college students, reached a market audience of 50 million people within just two years. In 2010 Facebook had more than 750 million active users, and more than 375 million users were logging on at least once each day. It has now moved into business applications, with business and interest groups having Facebook pages. It is also increasingly more common to use Facebook as the venue for organizing and conducting conferences.

Why are these communication innovations important? Beginning with text messaging, we need to consider the shift in grammar, syntax, and spelling that pervades these communications. If we consider the proliferation of videos on YouTube, it is important to see how effective different presentation forms of the same information can be. Similarly, Facebook presents even more challenges as it merges registers — here professional and personal communications can exist side-by-side.

One prominent example of incorporating new technologies into measures of communication was developed for PISA 2009. PISA's Electronic Reading Assessment simulated reading in a web environment. In many ways, this step forward represents not only migration to newer innovative assessment items but also a first step in transforming assessments to more authentic and up-to-date tasks.

Collaboration and Teamwork

Operational definitions of collaboration are provided in Table 2.6. Collaboration presents a different set of challenges for large-scale assessments. At the most basic, school level assessments are focused on getting measures of individual performance. Consequently, when faced with a collaborative task, the most important question is how to assign credit to each member of the group, as well as how to account for differences across groups that may bias a given student's performance. This becomes an even larger issue within international assessments where cultural boundaries are crossed. For example, ALL researched the potential for measuring teamwork. While the designers could generate teamwork tasks, at that time accounting for cultural differences became an insurmountable obstacle.

Several important research initiatives have worked on getting measures of individual performance that address key components of collaboration and measurement (Laurillard 2009). For example, Çakir et al. (2009) have shown how group participants, in order to collaborate effectively in group discourse on a topic like mathematical patterns, must organize their activities in ways that share the significance of their utterances, inscriptions, and behaviors. Their analysis reveals methods by which the group co-constructs meaningful inscriptions in the interaction spaces of the collaborative environment. The integration of graphical, narrative, and symbolic semiotic modalities facilitates joint problem solving. It allows group members to invoke and operate with multiple realizations of their mathematical artifacts, a characteristic of deep learning of mathematics. Other research shows how engaging in reflective activities in interaction, such as explaining, justifying, and evaluating problem solutions, collaboratively can potentially be productive for learning (Baker and Lund 1997). Several studies have also shown how taking part in collaborative inquiry toward advancing a shared knowledge object can serve as a means to facilitate the development of metaskills.

Table 2.6 Ways of working – collaboration, teamwork

Knowledge	Skills	Attitudes/values/ethics
Interact effectively with others • Know when it is appropriate to listen and when to speak *Work effectively in diverse teams* • Know and recognize the individual roles of a successful team and know own strengths and weaknesses, and recognizing and accepting them in others *Manage projects* • Know how to plan, set, and meet goals and to monitor and re-plan in the light of unforeseen developments	*Interact effectively with others* • Speak with clarity and awareness of audience and purpose. Listen with care, patience, and honesty • Conduct themselves in a respectable, professional manner *Work effectively in diverse teams* • Leverage social and cultural differences to create new ideas and increase both innovation and quality of work *Manage projects* • Prioritize, plan, and manage work to achieve the intended group result *Guide and lead others* • Use interpersonal and problem-solving skills to influence and guide others toward a goal • Leverage strengths of others to accomplish a common goal • Inspire others to reach their very best via example and selflessness • Demonstrate integrity and ethical behavior in using influence and power	*Interact effectively with others* • Know when it is appropriate to listen and when to speak • Conduct themselves in a respectable, professional manner *Work effectively in diverse teams* • Show respect for cultural differences and be prepared to work effectively with people from a range of social and cultural backgrounds • Respond open-mindedly to different ideas and values *Manage projects* • Persevere to achieve goals, even in the face of obstacles and competing pressures *Be responsible to others* • Act responsibly with the interests of the larger community in mind

Two further lines of research are pertinent to including collaborative work in large-scale assessments. The first line of research begins with the idea of a simulation where one respondent interacts with pre-programmed virtual partners. The drawback here is the current lack of theoretical understandings of how collaborators would interact in this environment. The second line of research is best exemplified by group tasks where evidence of interaction patterns and self-reflections are captured. Research into how to rate these interactions would lead to a rubric that might either be criterion-referenced or be normed according to country, nationality, socioeconomic status, or other differentiating group characteristics. In conjunction with the product scores, it would be possible to generate a collaboration scale on the basis of such research.

It has been observed that as employers, we most often base our staff recruitment decisions on formal, school, and college-based qualifications, using these as a measure of an applicant's potential to operate well within our organizations. However, we make decisions to fire people on the basis of their team-working skills, their collaborative styles, and their approach to work. These are the skills that matter most to us as employers, and it is in these areas that employers have for many years looked to occupational psychologists for support. There are a large number of psychological profiling measures, most of which seek to provide a prose summary of the interpersonal styles of working likely to be adopted by an individual. These profile measures attempt to score, for example, the extent to which an individual might seek help, might use discussion and dialogue to move matters forward, or might be an effective solver of open-ended and ill-defined problems. SHL provide assessments such as OPQ and 16PF, which are conducted online and are widely used by employers. The OPQ assessments seek to measure likely behaviors in three areas: Relationships with People, Thinking Style, and Feeling and Emotions. For example, in measuring Feeling and Emotions, OPQ gauges the extent to which an individual is *relaxed, worrying, tough minded, optimistic, trusting,* and *emotionally controlled.* Similarly, OPQ measures a dimension called Influence and gauges the extent to which an individual is *persuasive, controlling, outspoken,* and *independent minded.* These – and other measures, such as Belbin's team styles – provide considerable overlap with the skills domain that interests twenty-first century educators and could well provide useful examples of the ways in which it is possible to assess students' ways of working.

Tools for Working

The newest set of skills is combined in this grouping of tools for working. These skills, information literacy and ICT literacy, are the future and mark a major shift that is likely to be as important as the invention of the printing press. Friedman (2007) describes four stages in the growing importance of ICT. He identifies four "flatteners" that are making it possible for individuals to compete, connect, and collaborate in world markets:

- The introduction of personal computers that allowed anyone to author his/her own content in digital form that could then be manipulated and dispatched.
- The juxtaposition of the invention of the browser by Netscape that brought the internet to life resulting in the proliferation of websites and the overinvestment into fiber optic cable that has wired the world. NTT Japan has successfully tested a fiber optic cable that pushes 14 trillion bits per second that roughly equals 2,660 CDs or 20 million phone calls every second.
- The development of transmission protocols that made it possible for everyone's computer and software to be interoperable. Consequently, everyone could become a collaborator.

- The expansion of the transmission protocols so that individuals could easily upload as well as download. For example, when the world was round, individuals could download vast amounts of information in digital formats that they could easily access and manipulate. But, in the flat world, the key is the individual's ability to upload. This has given rise to open-source courseware, blogs, and Wikipedia, to name only a few examples.

To paint a picture of how important it is to be truly literate in the use of these tools, consider that it is estimated that a week's worth of the New York Times contains more information than a person was likely to come across in a lifetime in the eighteenth century. Moreover, it was estimated that four exabytes (4.0×10^{19}) of unique information was be generated in 2010 – more than that the previous 5,000 years put together. In light of this information explosion, the coming generations must have the skills to access and evaluate new information efficiently so they can effectively utilize all that is available and relevant to their tasks at hand. One of the ways that they will manage this information explosion is through skilled use of ICT. Even now the use of ICT is growing. It has been reported that there are 31 billion searches on Google every month, up from 2.7 billion in 2006. To use Google, one must effectively use the internet. To accommodate the use of the internet, we have seen an explosion in the number of internet devices. In 1984, the number was 1,000, by 1992 it was 1,000,000, and in 2008 it had reached 1,000,000,000.

Information Literacy

Information literacy includes research on sources, evidence, biases, etc. Operational definitions of information literacy are provided in Table 2.7. These are clearly increasingly important skills.

The future consequences of recent developments in our societies due to globalization, networking (Castells 1996), and the impact of ICT are spawning a set of new studies. Hull and Schultz (2002) and Burbules and Silberman-Keller (2006) are examples of how such developments change conceptions of formal and informal learning and what some term distributed or networked expertise (Hakkarainen et al. 2004). Measurement procedures or indicators are still not clear with regard to these more future-oriented skills. For example, the *ImpaCT2* concept mapping data from the UK strongly suggests that there is a mismatch between conventional national tests, which focus on pre-specified knowledge and concepts, and the wider range of knowledge that students are acquiring by carrying out new kinds of activities with ICT at home (Somekh and Mavers 2003). By using concept maps and children's drawings of computers in their everyday environments, the research generates strong indication of children's rich conceptualization of technology and its role in their world for purposes of communication, entertainment, or accessing information. It shows that most children acquire practical skills in using computers that are not part of the assessment processes that they meet in schools. Some research has shown that students who are active computer users consistently underperform on paper-based tests (Russell and Haney 2000).

Table 2.7 Tools for working – information literacy

Knowledge	Skills	Attitudes/values/ethics
Access and evaluate information • Access information efficiently (time) and effectively (sources) • Evaluate information critically and competently *Use and manage information* • Use information accurately and creatively for the issue or problem at hand • Manage the flow of information from a wide variety of sources • Apply a fundamental understanding of the ethical/legal issues surrounding the access and use of information • Basic understanding of the reliability and validity of the information available (accessibility/acceptability) and awareness of the need to respect ethical principles in the interactive use of IST *Apply technology effectively* • Use technology as a tool to research, organize, evaluate, and communicate information • Use digital technologies (computers, PDAs, media players, GPS, etc.), communication/networking tools, and social networks appropriately to access, manage, integrate, evaluate, and create information to successfully function in a knowledge economy	*Access and evaluate information* • Ability to search, collect, and process (create, organize, and distinguish relevant from irrelevant, subjective from objective, real from virtual) electronic information, data, and concepts and to use them in a systematic way *Use and manage information* • Ability to use appropriate aids, presentations, graphs, charts and maps to produce, present, or understand complex information • Ability to access and search a range of information media including the printed word, video, and websites and to use internet-based services such as discussion fora and email • Ability to use information to support critical thinking, creativity, and innovation in different contexts at home, leisure, and work • Ability to search, collect, and process written information, data, and concepts in order to use them in study and to organize knowledge in a systematic way; Ability to distinguish, in listening, speaking, reading, and writing, relevant from irrelevant information	*Access and evaluate information* • Propensity to use information to work autonomously and in teams; critical and reflective attitude in the assessment of available information *Use and manage information* • Positive attitude and sensitivity to safe and responsible use of the internet, including privacy issues and cultural differences • Interest in using information to broaden horizons by taking part in communities and networks for cultural, social and professional purposes

ICT Literacy

EU countries, both on a regional and national level, and other countries around the world are in the process of developing a framework and indicators to better grasp the impact of technology in education and what we should be looking for in assessing students' learning using ICT. Frameworks are being developed in Norway (see http://europa.eu/rapid/pressReleasesAction.do?reference=IP/09/1244), Norway (see Erstad, 2006), and Australia (see Ainley et al. 2006). According to the Summit on Twenty-first Century Literacy in Berlin in 2002 (Clift 2002), new approaches stress the abilities to use information and knowledge that extend beyond the traditional base of reading, writing, and mathematics, which has been termed *digital literacy* or *ICT literacy*. Operational definitions of information literacy are provided in Table 2.8.

In 2001, the Educational Testing Service (ETS) in the US assembled a panel for the purpose of developing a workable framework for ICT literacy. The outcome was the report *Digital transformation: A framework for ICT literacy* (International ICT Literacy Panel 2002). Based on this framework, shown in Table 2.9, one can define ICT literacy as "the ability of individuals to use ICT appropriately to access, manage and evaluate information, develop new understandings, and communicate with others in order to participate effectively in society." (Ainley et al. 2005) Different indicators of digital/ICT literacy can be proposed (Erstad 2010).

In line with this perspective, some agencies have developed performance assessment tasks of "ICT Literacy", indicating that ICT is changing our view on what is being assessed and how tasks are developed using different digital tools. One example is the tasks developed by the International Society for Technology in Education (ISTE) called *National Educational Technology Standards* (http://www.iste.org/standards.aspx), which are designed to assess how skilful students, teachers, and administrators are in using ICT.

In 2000, England's Department for Education commissioned the development of an innovative test of 14-year-old students' ICT skills. David Blunkett, at the time Secretary of State for Education, described his vision for education and attainment in the twenty-first century. He spoke of raising expectations of student capabilities. He also announced the development of a new type of online test of ICT, which would assess the ICT skills students need in the twenty-first century. These developed assessments are outlined in Fig. 2.3.

Development activity for the 14-year-old's test of ICT began in 2001. The original planned date for full roll-out and implementation was May 2009. In the event – and for a whole range of reasons – the original vision for the ICT tests was never realized. The test activities that were developed have been redesigned as standalone skills assessments that teachers in accredited schools can download and use informally to support their teacher assessment.

In Australia, a tool has been developed with a sample of students from grade 6 and grade 10 to validate and refine a progress map that identifies a progression of ICT literacy. The ICT literacy construct is described using three "strands": working with information, creating and sharing information, and using ICT responsibly. Students carrying out authentic tasks in authentic contexts are seen as fundamental to the design of the Australian National ICT Literacy Assessment Instrument (Ainley et al. 2005). The instrument evaluates six key processes: accessing information (identifying information requirements and knowing how to find and retrieve information); managing information (organizing and storing information for retrieval and reuse); evaluating (reflecting on the processes used to design and construct

Table 2.8 Tools for working – ICT literacy

Knowledge	Skills	Attitudes/values/ethics
Access and evaluate information and communication technology • Understanding of the main computer applications, including word processing, spreadsheets, databases, information storage and management • Awareness of the opportunities given by the use of Internet and communication via electronic media (e-mail, videoconferencing, other network tools) and the differences between the real and virtual world *Analyze media* • Understand both how and why media messages are constructed, and for what purposes • Examine how individuals interpret messages differently, how values and points of view are included or excluded, and how media can influence beliefs and behaviors • Understand the ethical/legal issues surrounding the access and use of media *Create media products* • Understand and know how to utilize the most appropriate media creation tools, characteristics, and conventions • Understand and know how to effectively utilize the most appropriate expressions and interpretations in diverse, multicultural environments	*Access and evaluate information and communication technology* • Access ICT efficiently (time) and effectively (sources) • Evaluate information and ICT tools critically and competently *Use and manage information* • Use ICT accurately and creatively for the issue or problem at hand • Manage the flow of information from a wide variety of sources • Apply a fundamental understanding of the ethical/legal issues surrounding the access and use of ICT and media • Employ knowledge and skills in the application of ICT and media to communicate, interrogate, present, and model *Create media products* • Utilize the most appropriate media creation tools, characteristics and conventions, expressions, and interpretations in diverse, multicultural environments *Apply technology effectively* • Use technology as a tool to research, organize, evaluate, and communicate information • Use digital technologies (computers, PDAs, media players, GPS, etc.), communication/networking tools, and social networks appropriately to access, manage, integrate, evaluate, and create information to successfully function in a knowledge economy • Apply a fundamental understanding of the ethical/legal issues surrounding the access and use of information technologies	*Access and evaluate information and communication technology* • Be open to new ideas, information, tools, and ways of working but evaluate information critically and competently *Use and manage information* • Use information accurately and creatively for the issue or problem at hand respecting confidentiality, privacy, and intellectual rights • Manage the flow of information from a wide variety of sources with sensitivity and openness to cultural and social differences • Examine how individuals interpret messages differently, how values and points of view are included or excluded, and how media can influence beliefs and behaviors *Apply and employ technology with honesty and integrity* • Use technology as a tool to research, organize, evaluate, and communicate information accurately and honestly with respect for sources and audience • Apply a fundamental understanding of the ethical/legal issues surrounding the access and use of information technologies

Table 2.9 Elaboration of key concepts of ICT literacy based on ETS framework

Category	Skills
Basic	Be able to open software, sort out and save information on the computer and other simple skills using the computer and software
Download	Be able to download different types of information from the internet
Search	Know about and how to get access to information
Navigate	Be able to orient oneself in digital networks, learning strategies in using the internet
Classify	Be able to organize information according to a certain classification scheme or genre
Integrate	Be able to compare and put together different types of information related to multimodal texts
Evaluate	Be able to check and evaluate if one has got the information one seeks to get from searching the internet. Be able to judge the quality, relevance, objectivity and usefulness of the information one has found. Critical evaluation of sources
Communicate	Be able to communicate information and express oneself through different meditational means
Cooperate	Be able to take part in net-based interactions of learning and take advantage of digital technology to cooperate and take part in networks
Create	Be able to produce and create different forms of information as multimodal texts, make web pages and so forth. Be able to develop something new by using specific tools and software. Remixing different existing texts into something new

ICT solutions and judgments regarding the integrity, relevance, and usefulness of information); developing new understandings (creating information and knowledge by synthesizing, adapting, applying, designing, inventing, or authoring); communicating (exchanging information by sharing knowledge and creating information products to suit the audience, the context, and the medium); and using ICT appropriately (critical, reflective, and strategic ICT decisions and considering social, legal, and ethical issues) (Ainley et al. 2005). Preliminary results of the use of the instrument show highly reliable estimates of ICT ability.

There are also cases where an ICT assessment framework is linked to specific frameworks for subject domains in schools. Reporting on the initial outline of a U.S. project aiming at designing a Coordinated ICT Assessment Framework, Quellmalz and Kozma (2003) have developed a strategy to study ICT tools and skills as an integrated part of science and mathematics. The objective is to design innovative ICT performance assessments that could gather evidence of use of ICT strategies in science and mathematics.

Living in the World

Borrowing the title of Bob Dylan's song, to say that "the times they are a changin'" is a gross understatement when one considers how different living and working in the world will soon be. For example, the U.S. Department of Labor estimated that today's learner will have between ten and fourteen jobs by age 38. This reflects rapidly growing job mobility, with one in four workers having been with their current employer for less than a year, and one in two has been there less than 5 years. One might ask where these people are going as manufacturing and service industries move to places where there are abundant sources of cheap but sufficiently educated labor supplies. Essentially, people must learn to live not only in their town or country but also in the world in its entirety. As more and more people individually move in the twenty-first century to compete, connect, and collaborate, it is even more important that they understand all the aspects of citizenship. It is not enough to assume that what goes on in your own country is how it is or should be all over the globe. Hence, we have identified and group Citizenship, Life and Career, and Personal and Social Responsibility together as twenty-first century skills.

Citizenship, Global and Local

Citizenship as an educational objective is not new and has been part of curricula, especially in social studies. A central focus has been on knowledge about democratic processes. Citizenship as a competence, however, has been growing in importance, and implies certain challenges in measurement. Operational definitions of citizenship are provided as shown in Table 2.10.

Honey led a worldwide investigation into the use of twenty-first century assessments which investigated the existence and quality of assessments in key areas, including global awareness, concluding that "no measures currently exist that address students' understanding of global and international issues." (Ripley 2007, p. 5)

One example of a large-scale assessment of citizenship skills is the International Civic Education Study conducted by the International Association for the Evaluation of Educational Achievement (IEA). This research tested and surveyed nationally representative samples consisting of 90,000 14 year-old students in 28 countries, and 50,000 17 to 19 year-old students in 16 countries throughout 1999 and 2000.

The content domains covered in the instrument were identified through national case studies during 1996–1997 and included democracy, national identity, social cohesion and diversity. The engagement of youth in civil society was also a focus. Torney-Purta et al. (2001) reported the findings from these studies in the following terms:

- Students in most countries have an understanding of fundamental democratic values and institutions – but depth of understanding is a problem.
- Young people agree that good citizenship includes the obligation to vote.
- Students with the most civic knowledge are most likely to be open to participate in civic activities.
- Schools that model democratic practice are most effective in promoting civic knowledge and engagement.

Table 2.10 Living in the world – citizenship, local and global

Knowledge	Skills	Attitudes/values/ethics
• Knowledge of civil rights and the constitution of the home country, the scope of its government • Understand the roles and responsibilities of institutions relevant to the policy-making process at local, regional, national, and international level • Knowledge of key figures in local and national governments; political parties and their policies • Understand concepts such as democracy, citizenship, and the international declarations expressing them • Knowledge of the main events, trends, and agents of change in national and world history • Knowledge of the movements of peoples and cultures over time around the world	• Participation in community/neighborhood activities as well as in decision making at national and international levels; voting in elections • Ability to display solidarity by showing an interest in and helping to solve problems affecting the local or wider community • Ability to interface effectively with institutions in the public domain • Ability to profit from the opportunities given by the home country and international programs	• Sense of belonging to one's locality, country, and (one's part of) the world • Willingness to participate in democratic decision making at all levels • Disposition to volunteer and to participate in civic activities and support for social diversity and social cohesion • Readiness to respect the values and privacy of others with a propensity to react against antisocial behavior • Acceptance of the concept of human rights and equality; acceptance of equality between men and women • Appreciation and understanding of differences between value systems of different religious or ethnic groups • Critical reception of information from mass media

- Aside from voting, students are skeptical about traditional forms of political engagement, but many are open to other types of involvement in civic life.
- Students are drawn to television as their source of news.
- Patterns of trust in government-related institutions vary widely among countries.
- Gender differences are minimal with regard to civic knowledge but substantial in some attitudes.
- Teachers recognize the importance of civic education in preparing young people for citizenship.

The main survey has been replicated as the International Civic and Citizenship Education Study in which data have been gathered in 2008 and 2009 and from which the international report was released in June 2010 (Schulz et al. 2010).

The developments of the internet and Web 2.0 technologies have implications for the conception of citizenship as a competence. Jenkins (2006) says these developments create a "participatory culture." This challenges, both locally and globally, the understanding of citizenship, empowerment, and engagement as educational priorities. At the moment, no measures exist which assess these skills in online environments, even though the research literature on "young citizens online" has been growing in recent years (Loader 2007).

One example of how these skills are made relevant in new ways is the Junior Summit online community. This consisted of 3,062 adolescents representing 139 countries. The online forum culminated in the election of 100 delegates. Results from one study indicate "young online leaders do not adhere to adult leadership styles of contributing many ideas, sticking to task, and using powerful language. On the contrary, while the young people elected as delegates do contribute more, their linguistic style is likely to keep the goals and needs of the group as central, by referring to the group rather than to themselves and by synthesizing the posts of others rather than solely contributing their own ideas. Furthermore, both boy and girl leaders follow this pattern of interpersonal language use. These results reassure us that young people can be civically engaged and community minded, while indicating that these concepts themselves may change through contact with the next generation" (Cassell et al. 2006). In this sense, it also relates to the German term "Bildung" as an expression of how we use knowledge to act on our community and the world around us, that is, what it means to be literate in a society, or what also might be described as cultural competence as part of broader personal and social responsibility.

Life and Career

The management of life and career is included among the skills needed for living in the world. There is a long tradition of measurement of occupational preferences as one component for career guidance but no strong basis for building measures of skill in managing life and career. Suggestions for building operational definitions of this skill are provided in Table 2.11.

Personal and Social Responsibility

The exercise of personal and social responsibility is also included among the skills needed for living in the world. There are aspects of this skill in collaboration and teamwork, which is among the skills included among ways of working. Personal and social responsibility is taken to include cultural awareness and cultural competence. There is not a body of measurement literature on which to draw, but the scope intended is set out in the operational definitions offered in Table 2.12.

Challenges

The foregoing discussions have laid out principles for the assessment of twenty-first century skills, proposed ten skills, and given a sense of what they are and what measurements related to them might be built upon. That being said, there is still a very long row to hoe, as it is not enough to keep perpetuating static tasks within the assessments. Rather, to reflect the need for imagination to compete, connect, and collaborate, it is essential that transformative assessments be created. This cannot begin to happen without addressing some very critical challenges.

Table 2.11 Living in the world – life and career

Knowledge	Skills	Attitudes/values/ethics
Adapt to change • Be aware that the twenty-first century is a period of changing priorities in employment, opportunity, and expectations • Understand diverse views and beliefs, particularly in multicultural environments *Manage goals and time* • Understand models for long-, medium-, and short-term planning and balance tactical (short-term) and strategic (long-term) goals *Be self-directed learners* • Identify and plan for personal and professional development over time and in response to change and opportunity *Manage projects* • Set and meet goals, even in the face of obstacles and competing pressures • Prioritize, plan, and manage work to achieve the intended result	*Adapt to change* • *Operate in varied roles, jobs responsibilities, schedules, and contexts* *Be flexible* • Incorporate feedback effectively • Negotiate and balance diverse views and beliefs to reach workable solutions *Manage goals and time* • Set goals with tangible and intangible success criteria • Balance tactical (short-term) and strategic (long-term) goals • Utilize time and manage workload efficiently *Work independently* • Monitor, define, prioritize, and complete tasks without direct oversight *Interact effectively with others* • Know when it is appropriate to listen and when to speak *Work effectively in diverse teams* • Leverage social and cultural differences to create new ideas and increase both innovation and quality of work *Manage projects* • Set and meet goals, prioritize, plan, and manage work to achieve the intended result even in the face of obstacles and competing pressures *Guide and lead others* • Use interpersonal and problem solving skills to influence and guide others toward a goal • Leverage strengths of others to accomplish a common goal • Inspire others to reach their very best via example and selflessness • Demonstrate integrity and ethical behavior in using influence and power	*Adapt to change* • Be prepared to adapt to varied responsibilities, schedules, and contexts; recognize and accept the strengths of others • See opportunity, ambiguity and changing priorities *Be flexible* • Incorporate feedback and deal effectively with praise, setbacks, and criticism • Be willing to negotiate and balance diverse views to reach workable solutions *Manage goals and time* • Accept uncertainty and responsibility and self manage *Be self-directed learners* • Go beyond basic mastery to expand one's own learning • Demonstrate initiative to advance to a professional level • Demonstrate commitment to learning as a lifelong process • Reflect critically on past experiences for progress *Work effectively in diverse teams* • Conduct self in a respectable, professional manner • Respect cultural differences, work effectively with people from varied backgrounds • Respond open-mindedly to different ideas and values *Produce results* • Demonstrate ability to: – Work positively and ethically – Manage time and projects effectively – Multi-task – Be reliable and punctual – Present oneself professionally and with proper etiquette – Collaborate and cooperate effectively with teams – Be accountable for results *Be responsible to others* • Act responsibly with the interests of the larger community in mind

Table 2.12 Living in the world – personal and social responsibility

Knowledge	Skills	Attitudes/values/ethics
• Knowledge of the codes of conduct and manners generally accepted or promoted in different societies • Awareness of concepts of individual, group, society, and culture and the historical evolution of these concepts • Knowledge of how to maintain good health, hygiene, and nutrition for oneself and one's family • Knowledge of the intercultural dimension in their own and other societies	• Ability to communicate constructively in different social situations (tolerating the views and behavior of others; awareness of individual and collective responsibility) • Ability to create confidence and empathy in other individuals • Ability to express one's frustration in a constructive way (control of aggression and violence or self-destructive patterns of behavior) • Ability to maintain a degree of separation between the professional and personal spheres of life and to resist the transfer of professional conflict into personal domains • Awareness and understanding of national cultural identity in interaction with the cultural identity of the rest of the world; ability to see and understand the different viewpoints caused by diversity and contribute one's own views constructively • Ability to negotiate	• Showing interest in and respect for others • Willingness to overcome stereotypes and prejudices • Disposition to compromise • Integrity • Assertiveness

This section summarizes key challenges to assessing twenty-first century skills in ways that truly probe the skills of students and provide actionable data to improve education and assessments.

Using Models of Skill Development Based on Cognitive Research

The knowledge about acquisition of twenty-first century skills and their development is very limited. The developers of assessments do not yet know how to create practical assessments using even this partial knowledge effectively (Bennett and Gitomer 2009).

Transforming Psychometrics to Deal with New Kinds of Assessments

Psychometric advances are needed to deal with a dynamic context and differentiated tasks, such as tasks embedded in simulations and using visualization that may yield a number of acceptable (and unanticipated) responses. While traditional assessments are designed to yield one right or best response, transformative assessments should be able to account for divergent responses, while measuring student performance in such a way that reliability of measures is ensured.

Making Students' Thinking Visible

Assessments should reveal the kinds of conceptual strategies a student uses to solve a problem. This involves not only considering students' responses but also interpreting their behaviors that lead to these responses. Computers can log every keystroke made by a student and thus amass a huge amount of behavioral data. The challenge is to interpret the meaning of these data and link patterns of behavior to the quality of response. These associations could then illuminate students' thinking as they respond to various tasks.

That computers can score student responses to items effectively and efficiently is becoming a reality. This is certainly true of selected-response questions where there is a single right answer. It is also quite easy to apply partial credit models to selected-response items that have been designed to match theories of learning where not quite fully correct answers serve as the distracters. Constructed responses pose challenges for automated scoring.

The OECD's PIAAC provides a good example of movement forward in machine scoring of short constructed responses. Some of the assessment tasks in PIAAC were drawn from the International Adult Literacy Survey (IALS) and the ALL Survey where all answers were short constructed responses that needed to be coded by humans. By altering the response mode into either drop and drag or highlighting, the test developers converted the items into machine scorable items. In these examples, however, all the information necessary to answer these types of questions resides totally in the test stimuli. Although the respondent might have to connect information across parts of the test stimuli, creation of knowledge not already provided is not required.

Machine scoring of extended constructed responses is in its infancy. Models do exist in single languages and are based on the recognition of semantic networks within responses. In experimental situations, these machine-scoring models are not only as reliable as human scorers but often achieve higher levels of consistency than can be achieved across human raters (Ripley and Tafler 2009). Work has begun in earnest to expand these models to cross languages and may be available for international assessments in the foreseeable future (Ripley 2009).

Interpreting Assisted Performance

New scoring rules are needed to take into account prompting or scaffolding that may be necessary for some students. Ensuring accessibility for as many students as possible and customization of items for special needs students within the design of the assessment are critical.

Assessing Twenty-First Century Skills in Traditional Subjects

Where the aims and goals of twenty-first century learning are described in countries' frameworks, they are generally specified as being taught through, within and across the subjects. However, computers can facilitate the creation of micro-worlds for students to explore in order to discover hidden rules or relationships. Tools such as computer-based simulations can, in this way, give a more nuanced understanding of what students know and can do than traditional testing methods. New approaches stress the abilities to use information and knowledge that extend beyond the traditional base of reading, writing, and mathematics. However, research shows that students still tuned into the old test situation with correct answers rather than explanations and reasoning skills can have problems in adjusting their strategies and skills. Without highly valued assessments of twenty-first century aims or goals requiring their teaching, it is difficult to see when or how education systems will change significantly for the majority of learners.

Accounting for New Modes of Communication

To date, newer modes of communication have rarely been represented in large-scale assessments. There is a mismatch between the skills young people gain in their everyday cultures outside of schools and the instruction and assessment they meet in schools. Different skills such as creativity, problem solving, and critical thinking might be expressed in different ways using different modes and modalities, which ICT provides. In light of the developments described in the chapter, it is essential that the radical changes in communication, including visual ways of communicating and social networking, be represented in some of the tasks of twenty-first century large-scale assessments. The speed with which new technologies develop suggests that it might be better to assess whether students are capable of rapidly mastering a new tool or medium than whether they can use current technologies.

Including Collaboration and Teamwork

Traditional assessments are focused on measuring individual performance. Consequently, when faced with a collaborative task, the most important question is how to assign credit to each member of the group, as well as how to account for differences across groups that may bias a given student's performance. This issue arises whether students are asked to work in pre-assigned complementary roles or whether they are also being assessed on their skills in inventing ways to collaborate in an undefined situation. Questions on assigning individual performance as well as group ratings become even more salient for international assessments where cultural boundaries are crossed.

Including Local and Global Citizenship

The assessment of citizenship, empowerment, and engagement, both locally and globally, is underdeveloped. At this time, no measures exist that assess these skills in online environments, even though the research literature on "young citizens online" has been growing in recent years. For international assessments, cultural differences and sensitivities will add to the challenge of developing tasks valid across countries. Having students solve problems from multiple perspectives is one way to address the challenge of cultural differences.

Ensuring Validity and Accessibility

It is important to ensure validity of standards on which assessments are based; accessibility with respect to skills demands, content prerequisites, and familiarity with media or technology and an appropriate balance of content and intellectual demands of tasks.

These important attributes of any assessments will prove particularly challenging for the transformative assessments envisaged in this paper. Careful development and piloting of innovative tasks will be required, including scoring systems that ensure comparability of complex tasks. Fluidity studies with technology are important in devising tasks for which experience with technology does not predict performance. Also, complex tasks typically demand access to intellectual resources (e.g., a search engine). This needs to be factored into designing complex assessment tasks as envisaged for transformative assessments.

Considering Cost and Feasibility

Cost and feasibility are factors operating for any assessment but will be greatly exacerbated for the innovative and transformative assessments that are to address the kinds of twenty-first century skills discussed in this paper. For sophisticated online assessments, ensuring that schools have both the technical infrastructure needed and the controls for integrity of data collection is mandatory. These latter matters are considered in Chap. 4.

References

Ainley, J., Fraillon, J., & Freeman, C. (2005). *National Assessment Program: ICT literacy years 6 & 10 report.* Carlton South, Australia: The Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA).

Ainley, J., Pratt, D., & Hansen, A. (2006). Connecting engagement and focus in pedagogic task design. *British Educational Research Journal, 32*(1), 23–38.

Anderson, R. (2009, April). A plea for ‘21st Century Skills’ white paper to include social and civic values. Memorandum to Assessment and Teaching of 21st Century Skills Conference, San Diego, CA.

Baker, E. L. (2007). The end(s) of testing. *Educational Researcher, 36*(6), 309–317.

Baker, M. J., & Lund, K. (1997). Promoting reflective interactions in a computer-supported collaborative learning environment. *Journal of Computer Assisted Learning, 13*, 175–193.

Banaji, S., & Burn, A. (2007). *Rhetorics of creativity*. Commissioned by Creative Partnerships. Retrieved November 30, 2009 www.creative-partnerships.com/literaturereviews

Bell, A., Burkhardt, H., & Swan, M. (1992). Balanced assessment of mathematical performance. In R. Lesh & S. Lamon (Eds.), *Assessment of authentic performance in school mathematics*. Washington, DC: AAAS.

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment, 1*(1), 14–15.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment. In C. Wyatt-Smith & J. Cumming (Eds.), *Assessment issues of the 21st Century*. New York: Springer Publishing Company.

Bennett, R. E., Jenkins, F., Persky, H., & Weiss, A. (2003). Assessing complex problem solving performances. *Assessment in Education: Principles, Policy & Practice, 10*, 347–360.

Black, P., McCormick, R., James, M., & Pedder, D. (2006). Learning how to learn and assessment for learning: A theoretical inquiry. *Research Papers in Education, 21*(2), 119–132.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–71.

Boeijen, G., & Uijlings, P. (2004, July). *Exams of tomorrow: Use of computers in Dutch national science exams.* Paper presented at the GIREP Conference, Teaching and learning physics in new contexts, Ostrava, Czech Republic.

Buckingham, D., & Willett, R. (Eds.). (2006). *Digital generations: Children, young people, and new media*. Mahwah: Lawrence Erlbaum.

Burbules, N. C., & Silberman-Keller, D. (2006). *Learning in places: The informal education reader*. New York: Peter Lang.

Çakir, M. P., Zemel, A., & Stahl, G. (2009). The joint organization of interaction within a multimodal CSCL medium. *International Journal of Computer-Supported Collaborative Learning, 4*(2), 115–149.

Cassell, J., Huffaker, D., Ferriman, K., & Tversky, D. (2006). The language of online leadership: Gender and youth engagement on the Internet. *Developmental Psychology, 42*(3), 436–449.

Castells, M. (1996). *The rise of the network society* (The information age: Economy, society and culture, Vol. 1). Cambridge: Blackwell.

Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). *Washback in language testing: Research contexts and methods*. Mahwah: Lawrence Erlbaum Associates.

Clift, S. (2002). *21st literacy summit white paper*. Retrieved from www.mail-archive.com/do-wire@tc.umn.edu/msg00434.html

Deakin Crick, R. D., Broadfoot, P., & Claxton, G. (2004). Developing an effective lifelong learning inventory: The ELLI project. *Assessment in Education: Principles, Policy & Practice, 11*, 247–318.

Draper, S. W. (2009). Catalytic assessment: Understanding how MCQs and EVS can foster deep learning. *British Journal of Educational Technology, 40*(2), 285–293.

Ericsson, K. A. (2002). Attaining excellence through deliberate practice: Insights from the study of expert performance. In M. Ferrari (Ed.), *The pursuit of excellence through education* (pp. 21–55). Mahwah: Lawrence Erlbaum Associates.

Erstad, O. (2006). A new direction? Digital literacy, student participation and curriculum reform in Norway. *Education and Information Technologies, 11*(3–4), 415–429.

Erstad, O. (2008). Trajectories of remixing: Digital literacies, media production and schooling. In C. Lankshear & M. Knobel (Eds.), *Digital literacies: Concepts, policies and practices* (pp. 177–202). New York: Peter Lang.

Erstad, O. (2010). Conceptions of technology literacy and fluency. In *International encyclopedia of education* (3rd ed.). Oxford: Elsevier.

Facione, P.A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction* (The Delphi Report). Millbrae: California Academic Press.

Forster, M., & Masters, G. (2004). Bridging the conceptual gap between classroom assessment and system accountability. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education*. Chicago: University of Chicago Press.

Friedman, T. (2007). *The world is flat.* New York: Farrar, Straus and Giroux.

Gardner, J. (Ed.). (2006). *Assessment & learning*. London: Sage Publications.

Gee, J. P. (2007). *What video games have to teach us about learning and literacy* (2nd ed.). New York: Palgrave Macmillan.

Gick, M., & Holyoak, K. (1983). Scheme induction and analogical transfer. *Cognitive Psychology, 15*(1), 1–38.

Gipps, C., & Stobart, G. (2003). Alternative assessment. In T. Kellaghan & D. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 549–576). Dordrecht: Kluwer Academic Publishers.

Hakkarainen, K., Palonen, T., Paavola, S., & Lehtinen, E. (2004). *Communities of networked expertise: Professional and educational perspectives*. Amsterdam: Elsevier.

Harlen, W. (2006). The role of assessment in developing motivation for learning. In J. Gardner (Ed.), *Assessment & learning* (pp. 61–80). London: Sage Publications.

Harlen, W., & Deakin Crick, R. (2003). Testing and motivation for learning. *Assessment in Education: Principles, Policy & Practice, 10*, 169–208.

Herman, J. L. (2008). Accountability and assessment in the service of learning: Is public interest in K-12 education being served? In L. Shepard & K. Ryan (Eds.), *The future of testbased accountability*. New York: Taylor & Francis.

Herman, J. L., & Baker, E. L. (2005). Making benchmark testing work. *Educational Leadership, 63*(3), 48–55.

Herman, J. L., & Baker, E. L. (2009). Assessment policy: Making sense of the babel. In D. Plank, G. Sykes, & B. Schneider (Eds.), *AERA handbook on education policy*. Newbury Park: Sage Publications.

Hof, R. D. (2007, August 20). Facebook's new wrinkles: The 35-and-older crowd is discovering its potential as a business tool. *Business Week*. Retrieved from http://www.businessweek.com/magazine/content/07_34/b4047050.htm

Holyoak, K. J. (2005). Analogy. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 117–142). Cambridge: Cambridge University Press.

Hull, G., & Schultz, K. (2002). *School's out! Bridging out-of-school literacies with classroom practice*. New York: Teachers College Columbia University.

International ICT Literacy Panel. (2002). *Digital transformation: A framework for ICT literacy*. Princeton: Educational Testing Service.

Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York: New York University Press.

Johnson, M., & Green, S. (2004). *Online assessment: The impact of mode on student performance.* Paper presented at the British Educational Research Association Annual Conference, Manchester, UK.

Koretz, D., Broadfoot, P., & Wolf, A. (Eds.). (1998). *Assessment in Education: Principles, policy & practice* (Special issue: Portfolios and records of achievement). London: Taylor & Francis.

Kozma, R. B. (Ed.). (2003). *Technology, innovation, and educational change: A global perspective*. Eugene: International Society for the Evaluation of Educational Achievement.

Laurillard, D. (2009). The pedagogical challenges to collaborative technologies. *International Journal of Computer-Supported Collaborative Learning, 4*(1), 5–20.

Lee, E. Y. C., Chan, C. K. K., & van Aalst, J. (2006). Students assessing their own collaborative knowledge building. *International Journal of Computer-Supported Collaborative Learning, 1*(1).

Lessig, L. (2008). *Remix: Making art and commerce thrive in the hybrid economy*. New York: Penguin Press.

Lin, S. S. J., Liu, E. Z. F., & Yuan, S. M. (2001). Web-based peer assessment: Feedback for students with various thinking styles. *Journal of Computer Assisted Learning, 17*, 420–432.

Loader, B. (Ed.). (2007). *Young citizens in the digital age: Political engagement, young people and new media*. London: Routledge.

Loveless, A. (2007). *Creativity, technology and learning*. (*Update*.) Retrieved November 30, 2009 http://www.futurelab.org.uk/resources/publications-reports-articles/literature-reviews/Literature-Review382

McFarlane, A. (2001). Perspectives on the relationships between ICT and assessment. *Journal of Computer Assisted Learning, 17*, 227–234.

McFarlane, A. (2003). Assessment for the digital age. *Assessment in Education: Principles, Policy & Practice, 10*, 261–266.

Mercer, N., & Littleton, K. (2007). *Dialogue and the development of children's thinking*. London: Routledge.

National Center on Education and the Economy. (1998). New standards: Performance standards and assessments for the schools. Retrieved at http://www.ncee.org/store/products/index.jsp?setProtocol=true&stSection=1

National Research Council (NRC). (1996). *National science education standards*. Washington, DC: National Academy Press.

No Child Left Behind Act of 2001, United States Public Law 107–110.

Nunes, C. A. A., Nunes, M. M. R., & Davis, C. (2003). Assessing the inaccessible: Metacognition and attitudes. *Assessment in Education: Principles, Policy & Practice, 10*, 375–388.

O'Neil, H. F., Chuang, S., & Chung, G. K. W. K. (2003). Issues in the computer-based assessment of collaborative problem solving. *Assessment in Education: Principles, Policy & Practice, 10*, 361–374.

OECD. (2005). *Formative assessment: Improving learning in secondary classrooms*. Paris: OECD Publishing.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know*. Washington, DC: National Academy Press.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3*(6). Available from. http://www.jtla.org, 4–30

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning and Assessment, 2*(6).

Quellmalz, E. S., & Kozma, R. (2003). Designing assessments of learning with technology. *Assessment in Education: Principles, Policy & Practice, 10*, 389–408.

Quellmalz, E., Kreikemeier, P., DeBarger, A. H., & Haertel, G. (2007). A study of the alignment of the NAEP, TIMSS, and New Standards Science Assessments with the inquiry abilities in the National Science Education Standards. Presented at the Annual Meeting of the American Educational Research Association, April 9–13, Chicago, IL

Raikes, N., & Harding, R. (2003). The horseless carriage stage: Replacing conventional measures. *Assessment in Education: Principles, Policy & Practice, 10*, 267–278.

Ridgway, J., & McCusker, S. (2003). Using computers to assess new educational goals. *Assessment in Education: Principles, Policy & Practice, 10*(3), 309–328.

Ridgway, J., McCusker, S., & Pead, D. (2004). *Literature review of e-assessment (report 10)*. Bristol: Futurelab.

Ripley, M. (2007). *E-assessment: An update on research, policy and practice*. Bristol: Futurelab. Retrieved November 30, 2009 http://www.futurelab.org.uk/resources/publications-reports-articles/literature-reviews/Literature-Review204

Ripley, M. (2009). JISC case study: Automatic scoring of foreign language textual and spoken responses. Available at http://www.dur.ac.uk/smart.centre1/jiscdirectory/media/JISC%20Case%20Study%20-%20Languages%20-%20v2.0.pdf

Ripley, M., & Tafler, J. (2009). JISC case study: Short answer marking engines. Available at http://www.dur.ac.uk/smart.centre1/jiscdirectory/media/JISC%20Case%20Study%20-%20Short%20Text%20-%20v2.0.pdf

Rumpagaporn, M. W., & Darmawan, I.N. (2007). Student's critical thinking skills in a Thai ICT schools pilot project. *International Education Journal*, *8*(2), 125–132. Retrieved November 30, 2009 http://digital.library.adelaide.edu.au/dspace/handle/2440/44551

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20). Retrieved from http://epaa.asu.edu/epaa/v7n20

Russell, M., & Haney, W. (2000). Bridging the gap between testing and technology in schools. *Education Policy Analysis Archives, 8*(19). Retrieved from http://epaa.asu.edu/epaa/v8n19.html

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: A look into the future. *Assessment in Education: Principles, Policy & Practice, 10*, 279–294.

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy and technology. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences*. New York: Cambridge University Press.

Schulz, W., Ainley, J., Fraillon, J., Kerr, D., & Losito, B. (2010). *Initial Findings from the IEA International Civic and Citizenship Education Study*. Amsterdam: IEA.

Sefton-Green, J., & Sinker, R. (Eds.). (2000). *Evaluating creativity: Making and learning by young people*. London: Routledge.

Shepard, L. (2007). Formative assessment: Caveat emptor. In C. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 279–304). Mahwah: Lawrence Erlbaum Associates.

Shepard, L., Hammerness, K., Darling-Hammond, D., & Rust, R. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do*. Washington, DC: National Academy of Education.

Shephard, K. (2009). E is for exploration: Assessing hard-to-measure learning outcomes. *British Journal of Educational Technology, 40*(2), 386–398.

Somekh, B., & Mavers, D. (2003). Mapping learning potential: Students' conceptions of ICT in their world. *Assessment in Education: Principles, Policy & Practice, 10*, 409–420.

Sweller, J. (2003). Evolution of human cognitive architecture. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 43, pp. 215–266). San Diego: Academic.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review,* (34) 273–286.

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and education in twenty-eight countries: Civic knowledge and engagement at age fourteen*. Amsterdam: IEA.

Voogt, J., & Pelgrum, W. J. (2003). ICT and the curriculum. In R. B. Kozma (Ed.), *Technology, innovation, and educational change: A global perspective* (pp. 81–124). Eugene: International Society for Technology in Education.

Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching* (*Studies in Language Testing*, Vol. 22). Cambridge: Cambridge University Press.

Walton, S. (2005). *The eVIVA project: Using e-portfolios in the classroom*. BETT. Retrieved June 7, 2007, from www.qca.org.uk/downloads/10359_eviva_bett_2005.pdf

Wasson, B., Ludvigsen, S., & Hoppe, U. (Eds.). (2003). *Designing for change in networked learning environments: Proceedings of the International Conference on Computer Support for Collaborative Learning 2003* (Computer-Supported Collaborative Learning Series, Vol. 2). Dordrecht: Kluwer Academic Publishers.

Webb, N.L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Madison: National Institute for Science Education.

Wegerif, R., & Dawes, L. (2004). *Thinking and learning with ICT: Raising achievement in primary classrooms*. London: Routledge Falmer.

Whitelock, D., with contributions from Road, M., & Ripley, M. (2007). *Effective practice with e-Assessment*. The Joint Information Systems Committee (JISC), UK. Retrieved November 30, 2009 http://www.jisc.ac.uk/publications/documents/pub_eassesspracticeguide.aspx

Williams, J. B., & Wong, A. (2009). The efficacy of final examinations: A comparative study of closed-book, invigilated exams and open-book, open-web exams. *British Journal of Educational Technology, 40*(2), 227–236.

Wilson, M., & Sloane, K. (2000). From principles to practice: an embedded assessment system. *Applied Measurement in Education, 13*(2), 181–208.

Woodward, H., & Nanlohy, P. (2004). Digital portfolios in pre-service teacher education. *Assessment in Education: Principles, Policy & Practice, 11*, 167–178.

Chapter 3
Perspectives on Methodological Issues

Mark Wilson, Isaac Bejar, Kathleen Scalise, Jonathan Templin, Dylan Wiliam, and David Torres Irribarra

Abstract In this chapter the authors have surveyed the methodological perspectives seen as important for assessing twenty-first century skills. Some of those issues are specific to twenty-first century skills, but the majority would apply more generally to the assessment of other psychological and educational variables. The narrative of the paper initially follows the logic of assessment development, commencing by defining constructs to be assessed, designing tasks that can be used to generate informative student responses, coding/valuing of those responses, delivering the tasks and gathering the responses, and modeling the responses in accordance with the constructs. The paper continues with a survey of the strands of validity evidence that need to be established, and a discussion of specific issues that are prominent in this context, such as the need to resolve issues of generality versus contextual specificity; the relationships of classroom to large-scale assessments; and the possible roles for technological advances in assessing these skills. There is also a brief segment discussing some issues that arise with respect to specific types of variables involved in the assessment of twenty-first century skills. The chapter concludes with a listing of particular challenges that are regarded as being prominent at the time of

M. Wilson (✉) • D.T. Irribarra
University of California, Berkeley
e-mail: MarkW@berkeley.edu

I. Bejar
Educational Testing Service

K. Scalise
University of Oregon

J. Templin
University of Georgia

D. Wiliam
Institute of Education, University of London

P. Griffin et al. (eds.), *Assessment and Teaching of 21st Century Skills*,
DOI 10.1007/978-94-007-2324-5_3, © Springer Science+Business Media B.V. 2012

writing. There is an annexure that describes specific approaches to assessment design that are useful in the development of new assessments.

Perhaps one of the most important, yet often overlooked, choices in assessment is how results are to be presented to various types of stakeholders. This is of prime importance since decisions that will influence the future learning of test takers are based on these results. Reflecting on the kinds of assessment reports that we want to provide is an excellent way to start thinking about the challenges that we face in designing assessment structures to support the development of twenty-first century skills. There have been several efforts to create lists of such skills—indeed, a companion paper provides a perspective on a range of these,[1] some examples being: creativity and innovation, collaboration (teamwork), and information literacy. Why are assessment reports a good starting point? Because they encourage us to think about the topics that we want to assess, invite us to consider what kind of inferences we want to promote to users, and lead us to ponder what kind of evidence we should deem appropriate to support those inferences.

The kinds of reports that we aspire to provide will be directly useful in enhancing instruction by targeting the teaching of the skills being assessed. Ideally, we want these reports to provide timely and easily interpretable feedback to a wide variety of users, including students and teachers, parents and principals, administrative authorities, and the general public. Finally, we want these reports to be valid and reliable by adhering to high technical standards in the development of the assessments and the analysis of the data.

A brief look at some of these topics leads to questions that need to be addressed. A few of the issues we face are:

- The selection of the constructs to be evaluated: Are these skills defined as domain-general or closely associated with specific contexts or disciplines?
- The age span of the skills: Will they be confined to K12, higher education, or beyond?
- The level of analysis at which we want to provide feedback: for individuals, teams, classes, or large groups?
- The question of the universality or cultural specificity of the skills.

The answers to these and other questions will shape decisions about the characterization of the constructs to be assessed, the kinds of instruments that will be developed, and the level of information that will be gathered. Ultimately, these decisions will delineate the available evidence and so will constrain the kinds of inferences that can be supported and communicated to users.

It is for this reason that it is extremely important to ensure that the development of our assessments is guided by the kinds of inferences that we want to encourage.

In this chapter, we present an overview of the assessment design process. The first section addresses the role of evidentiary reasoning, as the starting point of a sound assessment. Sections Two through Six review the different steps involved in the

[1] We will not specify a comprehensive list of the 21st century skills here. That is provided in Chap. 2.

development of an assessment, respectively: (a) defining the constructs to be measured, (b) creating the tasks that will be used to elicit responses and performances, (c) assigning values (codes or scores) to the student responses to these tasks, (d) gathering and delivering the responses, and (e) the modeling and analysis of those responses. Section Seven summarizes the various elements involved in constructing a validity argument to support the claims that will be based on the collected data. Section Eight discusses three general issues that need to be addressed in the design of assessments for twenty-first century skills, namely, the relation between content and process, the interactions between classroom-based and large-scale assessments, and finally, the opportunities that technology offers in the construction of assessments. Section Nine reviews examples of measures that can help visualize potential forms of assessments. A final section summarizes the issues and open challenges raised in the previous sections.

Inferences, Evidence, and Validity

As Mislevy et al. (2003a) have pointed out, assessment is a special kind of *evidentiary reasoning* in which evidence—defined as data that increases or decreases the likelihood of the acceptance of a claim (Schum 1987)—is used to support particular kinds of claims.

Since assessments are designed to support inferences, it is logical to begin with the inferences that are to be made and to work backwards from there. This is one of the central features of the evidence-centered approach to the design of education assessments (Mislevy et al. 2003a).

In early work on assessment in the first half of the twentieth century, it was assumed that inferences would be made with respect to a well-defined universe of content, for example, the 81 multiplication facts from 2×2 to 10×10. By sampling the competence of students on a random sample of these 81 facts, the laws of statistical inference could be used to estimate the proportion of these facts known by each individual, together with the precision of these estimates. However, it quickly became clear that for most of the inferences being sought, no such universe could be defined with sufficient completeness or accuracy and neither would it fit in with modern thinking about the development of student understanding.

Where the inferences were related to a criterion, such as performance in a subject at some future time, then evidence for the validity of the inferences could be derived from measures of correlation between the predictor and the criterion (Guilford 1946), and this led to the view of many in the assessment field during the 1950s and 1960s that predictive validity (and its variants) was the most important form of validity evidence. However, such approaches still left a large number of assessment situations without an adequate theoretical basis.

To address this, Cronbach and Meehl (1955) proposed that construct validity could be used for cases in which there was no easily defined universe of generalization and no sufficiently robust predictor–criterion relationships. Over the following 30 years or so, the idea that construct-based inferences should be at the heart of

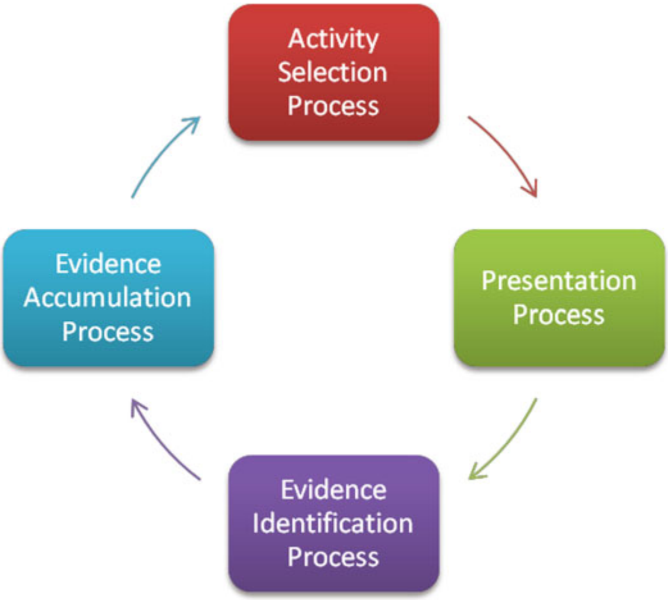

Activity Selection Process
Presentation Process
Evidence Identification Process
Evidence Accumulation Process

Fig. 3.1 The four-process architecture (Almond et al. 2003)

validity arguments became generally accepted, at least within the measurement community. This is why Messick (1995) has suggested that all assessment should be construct-referenced.

The starting point for the assessment of twenty-first century skills, therefore, must be an adequate construct definition, meaning one that defines the equivalence class of tasks for which successful performance will be taken as evidence of the presence of the construct (to a certain extent) and unsuccessful performance as evidence of its lack (to a certain extent).

Once the construct has been clarified, subsequent steps may be described in terms of the four-process architecture (see Fig. 3.1) proposed by Almond et al. (2003), in which tasks are *selected* on the basis of their relevance to the construct of interest and *presented* to learners. By engaging in the tasks, the learners generate evidence relevant to the *identified* construct of interest. Evidence from different sources (i.e., different tasks) is *accumulated*, which is then used to make inferences about the construct of interest.

One other thing to bear in mind is, as Messick (1989) has suggested, that a validity argument consists of not only showing that the evidence collected does support the intended inferences but also showing that plausible rival inferences are less warranted. This is where the specifications of the tasks are crucial, particularly in the context of twenty-first century skills. The collection of tasks presented to students must be designed and assembled in such a way that plausible rival interpretations—such as the fact that success might have been due to familiarity with the particular context rather than the underlying skill—are less warranted than the intended inferences.

Assessment Design Approaches

As the last section has indicated, the development of a good assessment system is rooted in the inferences that the system is intended to support—those inferences will frame and inform the development of the assessment. Successful development requires careful consideration of a series of elements, including (a) the definition and elaboration of the constructs that it is intended to measure; (b) the ways that those definitions guide the development and selection of the tasks or instruments that will be used to assess the constructs; and (c) ways of coding, classifying, or quantifying student responses, by assigning values to them (for instance, qualitative codes or quantitative scores) that relate back to the construct in meaningful ways.

We see these elements as common (in one form or another) to all assessments; they are taken into account in a variety of approaches to assessment design, for example, evidence-centered design (ECD; Mislevy et al. 2003b) and construct modeling (CM; Wilson 2005; Wilson and Sloane 2000) that attempt to systematize the assessment development process and provide a model for understanding the connections between these different elements. A summary of these two models can be found in the Annex. Because of the relevance of these elements to the development of assessments, they will be taken as the guiding structure for the next three sections.

Defining the Constructs

The importance of appropriate and meaningful definition of the skills to be assessed cannot be overstated. The success of any attempt to assess these skills will rely on these definitions and also on how they become elaborated as understanding evolves during the design and selection of the assessment instruments and activities. The same will apply during the appraisal of the products of the assessments.

The task of defining the different twenty-first century skills is not an easy one. As mentioned earlier, the definitions will need to address questions such as the unit of analysis (are they intended to reflect individuals, large groups, or both?); the age span of these skills (will they be confined to K12, higher education, or beyond?); whether the definitions are to be universal or susceptible to cultural differences; and whether the skills are to be defined as domain-general or closely associated with specific contexts or disciplines.

These are just some of the questions that need to be addressed during the definition of each skill, and the response to these questions will play a determining role in the delineation of the inferences that can be drawn from the assessment process. In other words, the definition of the constructs will determine the kind of information that will be collected, constraining the inferences that different stakeholders will be able to draw from the results of the assessment process.

Taking into account the overwhelming number of possible elements involved in each definition, where might we start to construct models of proficiency to serve as a solid base for assessment? Current literature in the field of educational assessment stresses that any measurement should be rooted in a robust cognitive theory[2] as well as a model of the learner that informs not only what counts as evidence of mastery but also the kinds of tasks that can be used to elicit them (NRC 2001). The Using Evidence framework provides an example of how a cognitive theory can be used as the basis of an assessment framework. It is described at the end of this section as a model of the use of evidence in scientific reasoning by students, teachers, and professional scientists and illustrates how cognitive theory may be linked to the different elements of an assessment system that are discussed throughout this report.

This leads to the key aspect that is emphasized in current learning theory, the need for a developmental[3] understanding of cognitive phenomena. This idea is clearly laid out in the NRC report, *How People Learn* (NRC 2000):

> The term "development" is critical to understanding the changes in children's conceptual growth. Cognitive changes do not result from mere accretion of information, but are due to processes of conceptual reorganization. (p. 234)

The elaboration of definitions rooted in a conception of cognitive growth confers meaning to the ideas of "improvement" and "learning" while describing and exemplifying what it means to become more proficient in each skill, and serves as a base for the definition of progress in each construct.

It is worth noting that a major aim of our emphasis on cognitive development is to help teachers build a common conception of progress, serving as a base for the coordination of instructional practice and assessment. That may require a substantial shift in view for some from a deficit and accretion model.

Structuring a Developmental Definition

When elaborating a developmental definition of a skill, the question remains about the characteristics that this kind of definition should have—what are the minimum elements that it should address? A recent report from the Center on Continuous Instructional Improvement (CCII) on the development of learning progressions, specific kinds of developmental perspectives, presents a summary of characteristics that are desirable when defining a developmental model of proficiency (CCII 2009):

- Learning targets,
- Progress variables,
- Levels of achievement,
- Learning performances.

[2] Although the emphasis in a cognitive perspective is often taken to be synonymous with information-processing views of cognition, this is by no means necessary. Alternative theoretical frameworks, such as sociocultural perspectives (Valsiner and Veer 2000) or embodied cognition approaches (Clark 1999), can be used to develop educational assessments.

[3] Note that the term "developmental" is not intended to imply that there is a biological inevitability to the process of development but that there are specific paths (not necessarily unique) that are seen as leading to more sophisticated learning.

These four elements are one possible guide to structure the developmental definition of each skill. We now examine each of them.

Learning Targets

Describing what the mastery of a given skill means is perhaps the first step in elaborating a developmental definition. A student who is fully accomplished in a skill can be seen as occupying a target point at the upper end of a progression defining previous stages of proficiency, with their corresponding performance level descriptors. The proficiency target at any given point in the teaching and learning trajectory might also differ from expert knowledge, while progressing toward it. Similarly, the proficiency targets could be clarified through the definition of "success criteria" on the construct, characterizing what success in the competencies for students in a given grade looks like. In any case, the point here is that clearly defining what mastery looks like is of the outmost importance.

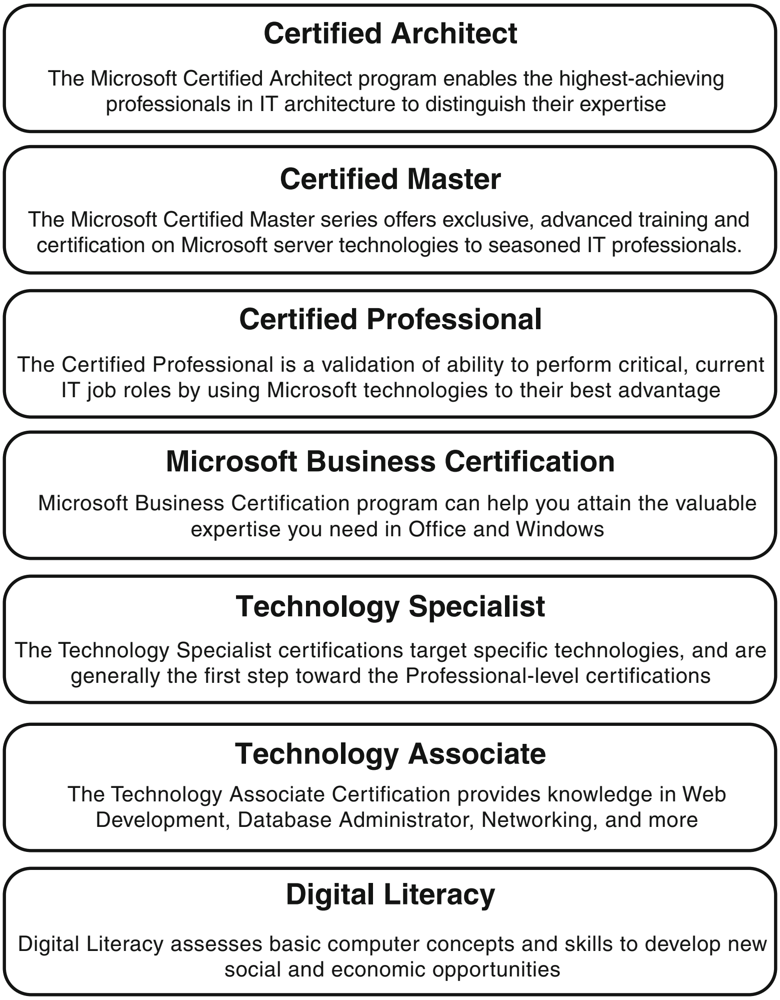
When defining learning targets, it is important to keep in mind that these target states exist within instructional contexts and so do not describe an inevitable outcome that would occur in the absence of instruction (Duncan and Hmelo-Silver 2009). In this sense, what constitutes mastery of a certain skill should be linked to curricular objectives and defined under the conditions of typical instruction.

An example of how learning targets can contribute to generating developmental definitions and delineating progress variables can be seen in the structure of Microsoft's certification program, discussed in the next section. In this case, we can see that defining the various learning targets at different levels of proficiency can convey the objectives of a curricular sequence. It is important to note that in the case of Microsoft's certification program, the progress variable is delineated at a very high level, hinting that the use of developmental progressions has potential in supporting the organization of long-term curricular sequences. At the same time, it is important to remember that each of the learning targets in this example has an important set of sublevels, with much more finely grained descriptions.

Learning Target: An Example from Microsoft Learning

The advantages of the idea of mapping progressions of proficiency are not restricted to school settings. Moreover, they can be a powerful and intuitive way of organizing different levels of competencies associated with different roles in professional settings. An example of how a progression can be developed in this context is offered by the structure of Microsoft's certification program presented in Fig. 3.2 (http://www.microsoft.com/learning/).

It is worth noticing that, although in this example the structure of the certification program can be easily understood as a learning progression, there is a subtle difference



Certified Architect

The Microsoft Certified Architect program enables the highest-achieving professionals in IT architecture to distinguish their expertise

Certified Master

The Microsoft Certified Master series offers exclusive, advanced training and certification on Microsoft server technologies to seasoned IT professionals.

Certified Professional

The Certified Professional is a validation of ability to perform critical, current IT job roles by using Microsoft technologies to their best advantage

Microsoft Business Certification

Microsoft Business Certification program can help you attain the valuable expertise you need in Office and Windows

Technology Specialist

The Technology Specialist certifications target specific technologies, and are generally the first step toward the Professional-level certifications

Technology Associate

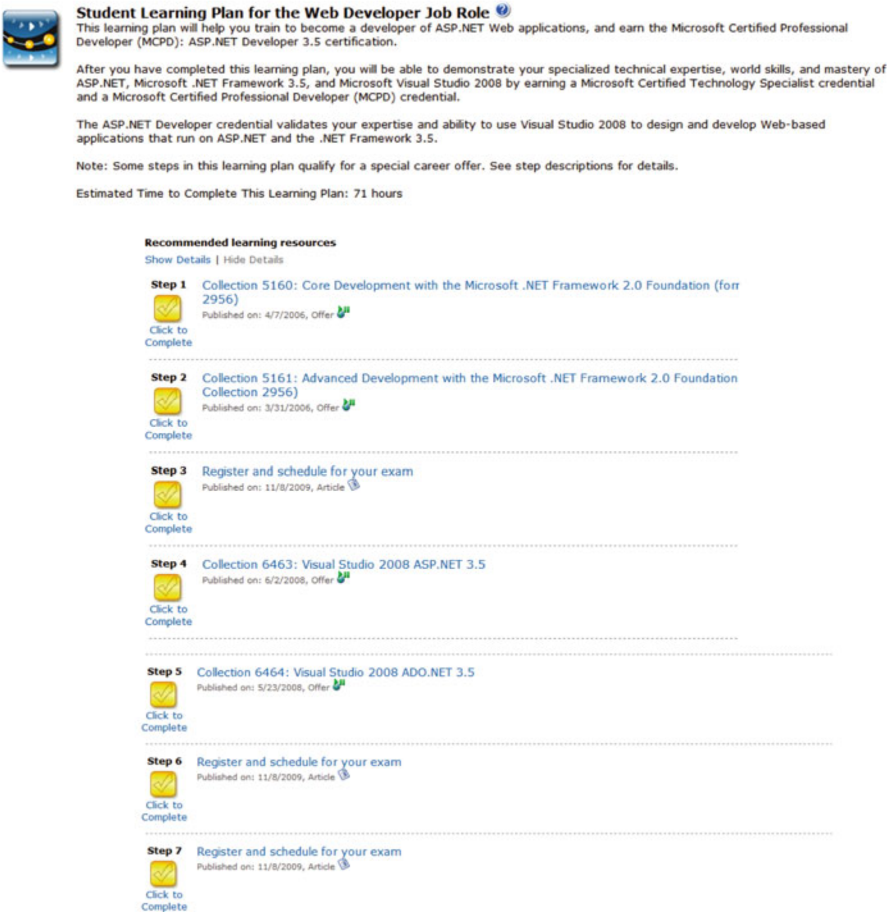
The Technology Associate Certification provides knowledge in Web Development, Database Administrator, Networking, and more

Digital Literacy

Digital Literacy assesses basic computer concepts and skills to develop new social and economic opportunities

Fig. 3.2 Structure of Microsoft's certification program

from the usual levels found in a typical academic setting. Within a school setting, there is a tendency for the lower levels of proficiency to represent misconceptions or incomplete preconceptions that will be overcome if the student successfully achieves mastery of the concept. In the case of this certification program, each level represents a target state of proficiency associated with a distinct role in an organization. This difference brings out an important possibility afforded by the creation of progressions as a basis for an assessment, namely, the possibility of organization within larger hierarchical frameworks. Another way to think about this issue is that the diagram presented in Fig. 3.2 does not represent seven levels of a single progression but seven smaller progressions stacked on top of one another. This understanding of the progression



Student Learning Plan for the Web Developer Job Role
This learning plan will help you train to become a developer of ASP.NET Web applications, and earn the Microsoft Certified Professional Developer (MCPD): ASP.NET Developer 3.5 certification.

After you have completed this learning plan, you will be able to demonstrate your specialized technical expertise, world skills, and mastery of ASP.NET, Microsoft .NET Framework 3.5, and Microsoft Visual Studio 2008 by earning a Microsoft Certified Technology Specialist credential and a Microsoft Certified Professional Developer (MCPD) credential.

The ASP.NET Developer credential validates your expertise and ability to use Visual Studio 2008 to design and develop Web-based applications that run on ASP.NET and the .NET Framework 3.5.

Note: Some steps in this learning plan qualify for a special career offer. See step descriptions for details.

Estimated Time to Complete This Learning Plan: 71 hours

Recommended learning resources
Show Details | Hide Details

Step 1 Collection 5160: Core Development with the Microsoft .NET Framework 2.0 Foundation (forr 2956)
Published on: 4/7/2006, Offer
Click to Complete

Step 2 Collection 5161: Advanced Development with the Microsoft .NET Framework 2.0 Foundation Collection 2956)
Published on: 3/31/2006, Offer
Click to Complete

Step 3 Register and schedule for your exam
Published on: 11/8/2009, Article
Click to Complete

Step 4 Collection 6463: Visual Studio 2008 ASP.NET 3.5
Published on: 6/2/2008, Offer
Click to Complete

Step 5 Collection 6464: Visual Studio 2008 ADO.NET 3.5
Published on: 5/23/2008, Offer
Click to Complete

Step 6 Register and schedule for your exam
Published on: 11/8/2009, Article
Click to Complete

Step 7 Register and schedule for your exam
Published on: 11/8/2009, Article
Click to Complete

Fig. 3.3 Example of a learning plan associated with a job role

seems intuitive in the context of professional development, reinforcing the idea that intermediate levels in a progression can be legitimate proficiency targets on their own. Moreover, depending on the extent of aggregation, it illustrates that an intermediate level can correspond to an entire progression in its own right.

In the case of Microsoft's certification program, this "nested" understanding fits well with the structure of their curriculum. Each one of these seven levels is defined by a set of target competencies for specific roles, and each role is associated with a structured collection of lectures that should lead to the achievement of those competencies. Figure 3.3 presents details of one of the learning plans for the role of Web developer. This is another example of how the progressions can serve as links connecting the structure of the curriculum with that of the assessment, where lessons are explicitly connected to both target proficiencies and assessment milestones (Microsoft 2009).

Progress Variables

The elaboration of learning targets will allow highlighting of the core themes of a domain; the themes will serve as the *central conceptual structures* (Case and Griffin 1990) or "big ideas" (Catley et al. 2005) that need to be modeled within each skill.

The notion of these central conceptual structures or themes is consistent with studies on expert–novice differences, which highlight how experts organize their knowledge according to major principles that reflect their deep understanding of a domain (National Research Council 2000).

The evolution of each of these themes can be represented as one or more progress variables that describe pathways that learners are likely to follow to progressively higher levels of performance and ultimately, for some, to mastery of a domain (CCII 2009). They can also help to explain how learning may proceed differently for different learners, depending on the strength of available theory and empirical evidence to support these findings.

It is worth clarifying that defining these pathways does not imply a single "correct" model of growth; it is important to recognize the remarkably different ways by which students can achieve higher levels of proficiency. Our ability to capture this diversity will depend to a certain extent both on the quality of our cognitive models (for interpreting this variation in substantive terms) and on the nature of our measurement models. The most critical element to keep in mind, however, is that making inferences at this level of detail about variations in individual developmental pathways will involve specific demands in terms of the quantity and specificity of the data to be collected.

Since these progress variables constitute the different elements that comprise each skill, they shed light on its dimensionality. Some skills may be appropriately defined in terms of a single theme, hence requiring only a single progress variable to characterize their development, while others may require more than one theme, increasing the need for a multidimensional model. An example of a progress variable that characterizes student responses in several dimensions is also presented later on when discussing the Using Evidence framework Brown et al. (2008, 2010a, 2010b). It allows the evaluation of students' scientific reasoning not only in terms of the correctness of their statements but also in terms of their complexity, validity, and precision, illustrating how progress variables can be used to capture different facets of complex processes. When considering situations where there is more than a single dimension, there are several approaches that build on this perspective and could help portray the increasing complexity and sophistication of each skill.

Measurement models, broadly conceived, can be considered to include multidimensional latent variable models, latent class models, and other models that might involve linear/nonlinear trajectories, transitive probabilities, time-series modeling, growth models, cognitive process models, or other methods. Multiple methodologies should be encouraged so as to balance the strengths and weaknesses of different techniques and to validate findings in this complex area of learning progressions.

Levels of Achievement

As mentioned in the previous section, each progress variable delineates a pathway (or pathways) that, based on a specific theory of a skill, characterizes the steps that learners may typically follow as they become more proficient (CCII 2009). Levels of achievement form one example of these different steps, describing the breadth and depth of the learner's understanding of the domain at a particular level of advancement (CCII 2009). It is important to keep in mind that the description of a level of achievement must "go beyond labels," fleshing out the details of the level of proficiency being described.

Learning Performances[4]

In the CCII report about learning progressions, learning performances are considered as:

> … the operational definitions of what children's understanding and skills would look like at each of these stages of progress, and … provide the specifications for the development of assessments and activities which would locate where students are in their progress. (CCII 2009, p. 15)

This term has been adopted by a number of researchers, e.g., Reiser (2002) and Perkins (1998), as well as by the NRC Reports "Systems for State Science Assessment" (NRC 2006) and "Taking Science to School" (NRC 2007). The idea is to provide a way of clarifying what is meant by a standard through describing links between the knowledge represented in the standard and what can be observed and thus assessed. Learning performances are a way of enlarging on the content standards by spelling out what one should be able to do to satisfy them. For example, within a science education context, learning performances lay out ways that students should be able to describe phenomena, use models to explain patterns in data, construct scientific explanations, or test hypotheses: Smith et al. (2006) summarized a set of observable performances that could provide indicators of understanding in science (see Fig. 3.4[5]).

As a concrete example, take the following standard, adapted from *Benchmarks for Science Literacy* (AAAS 1993, p. 124), about differential survival:

> [The student will understand that] Individual organisms with certain traits are more likely than others to survive and have offspring.

The standard refers to one of the major processes of evolution, the idea of "survival of the fittest." But it does not identify which skills and knowledge might be called for in working to attain it. In contrast, Reiser et al. (2003, p. 10) expand this single standard into three related learning performances:

[4] The following section was adapted from the NRC 2006 report *Systems for state science assessment* edited by Wilson & Bertenthal.

[5] Note that this is only a partial list of what is in the original.

Some of the key practices that are enabled by scientific knowledge include the following:

- **Representing** data and interpreting representations. Representing data involves using tables and graphs to organize and display information both qualitatively and quantitatively. Interpreting representations involves being able to use legends and other information to infer what something stands for or what a particular pattern means. For example, a student could construct a table to show the properties of different materials or a graph that relates changes in object volume to object weight. Conversely, a student could interpret a graph to infer which size object was the heaviest or a straight line with positive slope to mean there was proportionality between variables.
- **Identifying and classifying**. Both identifying and classifying involve applying category knowledge to particular exemplars. In identifying, students may consider only one exemplar (Is this particular object made of wax?) whereas in classifying students are organizing sets of exemplars. For example, they could sort items by whether they are matter or not matter; by whether they are solid, liquid, or gas; or by kind of substance.
- Measuring. Measuring is a simple form of mathematical modeling: comparing an item to a standard unit and analyzing a dimension as an iterative sum of units that cover the measurement space.
- **Ordering/comparing along a dimension**. Ordering involves going beyond simple categorization (e.g., heavy vs. light) to conceptualizing a continuous dimension. For example, students could sort samples according to weight, volume, temperature, hardness, or density.
- **Designing and conducting investigations**. Designing an investigation includes identifying and specifying what variables need to be manipulated, measured, and controlled; constructing hypotheses that specify the relationship between variables; constructing/developing procedures that allow them to explore their hypotheses; and determining how often the data will be collected and what type of observations will be made. Conducting an investigation includes a range of activities—gathering the equipment, assembling the apparatus, making charts and tables, following through on procedures, and making qualitative or quantitative observations.
- **Constructing evidence-based explanations**. Constructing explanations involves using scientific theories, models, and principles along with evidence to build explanations of phenomena; it also entails ruling out alternative hypotheses.
- **Analyzing and interpreting data.** In analyzing and interpreting data, students make sense of data by answering the questions: "What do the data we collected mean?" "How do these data help me answer my question?" Interpreting and analyzing can include transforming the data by going from a data table to a graph, or by calculating another factor and finding patterns in the data.
- **Evaluating/reflecting/making an argument**. Evaluate data: Do these data support this claim? Are these data reliable? Evaluate measurement: Is the following an example of good or bad measurement? Evaluate a model: Could this model represent a liquid? Revise a model: Given a model for gas, how would one modify it to represent a solid? Compare and evaluate models: How well does a given model account for a phenomenon? Does this model "obey" the "axioms" of the theory?

Fig. 3.4 Examples of evidence of understanding in science (From Smith et al. 2004)

> Students *identify and represent mathematically* the variation on a trait in a population.
> Students *hypothesize* the function a trait may serve and *explain* how some variations of the trait are advantageous in the environment.
> Students *predict, using evidence*, how the variation on the trait will affect the likelihood that individuals in the population will survive an environmental stress.

Reiser et al. (2003) advance the claim that this extension of the standard makes it more useful because it defines the skills and knowledge that students need in order to master the standard and therefore better identifies the construct (or learning progression) of which the standard is a part. For example, by explaining that students are expected to characterize variation mathematically, the extension makes clear the importance of specific mathematical concepts, such as distribution. Without this extension, the requirement for this important detail may have not been clear to a test developer and hence could have been left out of the test.

Assessment of Progressions

The four elements discussed above, learning targets, progress variables, levels of achievement, and learning performances, will allow us to formulate the different constructs in terms of *learning progressions*. The concept of a learning progression can be understood as one of the more recent incarnations of a familiar notion in the fields of cognition and development (NRC 2006), namely, that students can become more proficient in a domain by following trajectories of increasing complexity with support from appropriately structured learning contexts.

In discussing learning progressions, Duncan and Hmelo-Silver (2009) point out that the idea of learning progression is akin to earlier theoretical developments focused on development and deepening of knowledge over time, such as the concept of "bandwidths of competence" (Brown and Reeves 1987), and cognitively guided instruction (CGI; Carpenter and Lehrer 1999).

Learning progressions describe pathways that learners are likely to follow toward the mastery of a domain, providing models that on the one hand allow empirical exploration of their validity (CCII 2009) and on the other hand provide a practical tool for organizing instructional activities.

Notably, the educational usefulness of these models rests on determining a student's position along a learning progression. So, for a measurement approach to support a learning progression, its *assessment design* is crucial for its study and use.

According to a recent National Research Council report (NRC 2007), learning progressions are:

> …descriptions of the successively more sophisticated ways of thinking about an important domain of knowledge and practice that can follow one another as children learn about and investigate a topic over a broad span of time. They are crucially dependent on instructional practices if they are to occur. (p. 219)

Brown et al. (2008, 2010a, 2010b) propose the Using Evidence (UE) framework as a model of the use of evidence in scientific reasoning by students, teachers, and professional scientists. The main purpose of the model is to help researchers and practitioners identify the structure of scientific argumentation in student work and classroom discourse (Brown et al. 2008, 2010a).

Defining the Constructs—Example: The Using Evidence Framework

The UE framework (Brown et al. 2008, 2010a, 2010b) offers a theoretical perspective of scientific reasoning that can serve as the basis to a wide range of assessment tools including written products or classroom discussions. A diagram of the UE framework is presented in Fig. 3.5 (Brown et al. 2010a).

The key elements of the UE framework as described by Brown et al. (2010a) are:

- *The claims*: statements about outcomes in the form of predictions (e.g., "this box will sink"), observations (e.g., "this box sank"), or conclusions (e.g., "this box sinks") about the circumstances defined by the premise.
- *The premises*: statements that describe specific circumstances; in classroom contexts, premises usually identify objects and relevant features (e.g., "this box is heavy").

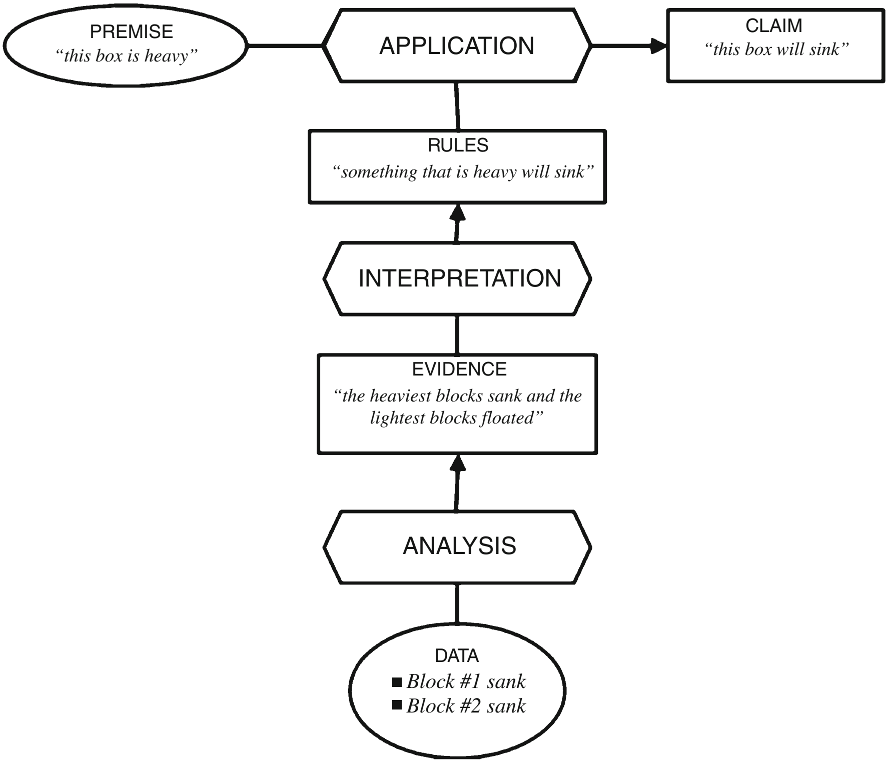


Fig. 3.5 The Using Evidence framework (Brown et al. 2010a)

- *The rules*: connections that indicate how the claim follows from the premise by stating general relationships. These relations are expected to hold even in contexts not previously observed (e.g., "something that is heavy will sink").
- *The application*: is the process that connects the rules to the specific circumstances described in the premise, establishing the probability or necessity of the claim. Depending on the complexity of the circumstances, it can vary from informal deductive logic to complex systems of analysis (e.g., "this box is heavy, heavy things sink, therefore this box will sink.").

Brown et al. (2010a) indicate that the UE framework *"describes scientific reasoning as a two-step process in which a uniquely scientific approach to gathering and interpreting data results in rules (theories, laws, etc.) that are applied within a general framework of argumentation in which claims are justified.*" (p. 133). In this framework *rules* play a central role in the scientific reasoning process, and are supported by the following elements (Brown et al. 2010a):

- *The evidence*: statements that describe observed relationships. (e.g., "the heaviest blocks sank and the lightest blocks floated" relates the weight with the behavior of the blocks). *Rules* are the product of the *interpretation* of evidence.

Table 3.1 Sample item prompts (Brown et al. 2010b)

Use the following information to answer Questions 3a and 3b
Here are some things that float in water:
A. A kitchen sponge
B. A plastic toy boat
C. An empty glass bottle
3a. What do these things have in common that causes them to float in water?
3b. Scientists require evidence to support their beliefs. Describe a specific thing you've seen, heard, or done that supports your belief that things float because of the reason you described in 3a

- *The data*: reports of observations (e.g., "Block #1 sank"), recollections (e.g., "my toy boat floats in my bathtub"), or thought experiments (e.g., "if I were to drop a tire in the ocean, it would float"). Statements of *evidence* are the product of the collection and *analysis* of these observations.

This framework allows different aspects of scientific reasoning to be selected as a focus for assessment and subsequent interpretation, and it serves as an example of how a cognitive model of a complex and dynamic process can be connected both to the generation of developmental hypotheses and the creation of rationales for evaluating students' responses. An example of one of the tasks that have been used by Brown et al. (2010a) in order to elicit students' reasoning on the topic of buoyancy is presented in Table 3.1.

Starting Point for a Developmental Progression

In developmental terms, the most important element of the UE framework is that it describes the state of proficiency that advanced students should achieve at the end of the instruction process. In this case, the authors of the model consider that a proficient response would contain elements of all five components of the model (premise, claim, rules, evidence, and data). At the same time, the model can be utilized to organize and describe the characteristics of the lower levels of proficiency. Broadly stated, the hypothesis is that lower proficiency levels will be expressed by incomplete arguments (Brown et al. 2010a). Figure 3.6 shows an example of a progression between three common incomplete argument structures that are hypothesized to constitute a hierarchy; it is important to note, however, that this is not a fully developed progression but only represents snapshots of "levels" that are common among students.

Another important aspect of the UE framework is that it allows a multidimensional understanding of the developmental progression, including the "correctness" of the statements, the sophistication of their structure, the precision of the responses, and their validity (Brown et al. 2010b). As an example, Table 3.2 summarizes the levels for two of these constructs that can be used to interpret and understand the student responses to the tasks (Brown et al. 2010b).

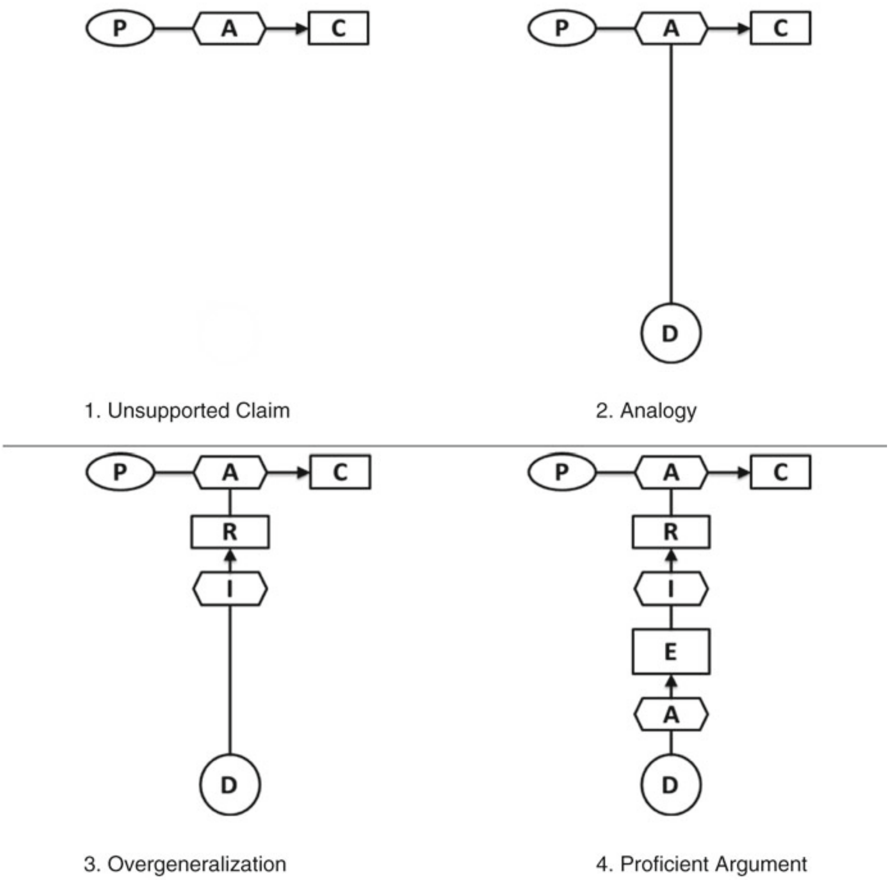


Fig. 3.6 Examples of levels in progression of quality of scientific argument. *1* Unsupported claim, *2* analogy, *3* overgeneralization, *4* proficient argument (Brown et al. 2010a)

Designing Tasks[6]

Once we have defined the construct, we need to be specific about the sort of performance that will convince an observer that the students have achieved mastery of the skills. Eventually, this will also need to be addressed by studies of validity questions ("How justified are we in drawing the intended conclusions from the assessment outcomes?") and reliability questions ("Are the responses consistent?").

[6] Some segments on the following section have been adapted from Wilson 2005.

Table 3.2 Validity and precision outcome spaces (Brown et al. 2010b)

Validity of the argument		Precision of the argument	
Response category	Description	Response category	Description
Fully valid	Entire conclusion follows from assumptions	Exact	Explicitly describes the exact value of properties
Partially valid	Part of conclusion follows from assumptions; rest of conclusion not warranted	Inexact	Implies the exact value of properties
Invalid	Conclusion is incorrectly based on assumptions	Vague	Describes the magnitude of properties
No link	Assumptions make it impossible to draw a conclusion	Indeterminate	States properties without magnitude

This section comprises four main topics: (1) the design of the assessment tasks and the way in which they can be organized in an overall taxonomy, (2) the valuation of the responses and performances obtained through the tasks in order to clarify the relation between the responses and the construct, (3) the challenges and opportunities raised by the assessment of twenty-first century skills, and (4) the issue of the different forms of delivery for the assessment tasks.

Our ability to use assessments in order to learn about students in any instructional context depends on our capacity to elicit products or actions that will provide information about the construct of interest. The quality of the tasks that we use to evoke this information about the progress variable is important because it will determine whether we consider these observable responses as valid evidence of the proficiency level of the student. It is important, therefore, to define in advance the type of evidence that is acceptable.

The creation and selection of tasks play an important role not only for the obvious reason that they will ultimately constitute the assessment but also because in many, if not most tasks, the construct itself will not be clearly defined until a large set of tasks has been developed and tried out with students. Simply stated, the design of the tasks helps clarify the construct that is being measured, bringing into focus any ambiguities or aspects that have not been well discerned. This is not to diminish the importance of clear, initial definition of the construct but rather to recognize the role of evidence in the initial design phase in sharpening and, when necessary, reshaping the definition.

The relationship of the task to the construct is important. Typically, the task is but one of many that could be used to measure the construct. Where one wishes to represent a wide range of contexts in an instrument, it is better to have more tasks rather than fewer, balancing this against the requirement to use item formats that are sufficiently complex to bring rich enough responses that will stand the sorts of interpretation that the measurer wishes to make of the measures. And both requirements need to be satisfied within the time and cost limitations of the measuring context.

Tasks can be characterized by their different amounts of prespecification—that is by the degree to which the possible outcomes of the instrument are structured before the instrument is administered to a respondent. The more that is prespecified, the less that has to be done after the response has been received.

Participant Observation

The item format with the lowest possible level of prespecification is one for which the developer has not yet formulated ANY of the item characteristics discussed above or even perhaps the construct itself, the very aim of the instrument. What is left here is the simple intent to observe. For some, this format may not even qualify as worthy of inclusion here—in that case, its inclusion should be considered as a device to define a lower end. This type of very diffuse instrumentation is exemplified by the *participant observation* technique (e.g., Ball 1985) common in anthropological studies. Another closely related technique is the "informal conversational interview" as described by Patton (1980):

> …The phenomenological interviewer wants to maintain maximum flexibility to be able to pursue information in whatever direction appears to be appropriate, depending on the information that emerges from observing a particular setting or from talking to one or more individuals in that setting. (pp. 198–199)

Not only is it the case that the measurer (i.e., in this case usually called the "participant observer") might not know the purpose of the observation but also "the persons being talked with might not even realize they are being interviewed" (Patton 1980, p. 198). The degree of prespecification of the participant observation item format is shown in the first row of Table 3.3, which emphasizes the progressive increase in prespecification as one moves from participant observation to fixed-response formats. It is not clear that one should consider a technique like participant observation as an example of an "instrument" at all. But it is included here because these techniques can be useful *within* an instrument design, and the techniques mark a useful starting point in thinking about the level of prespecification of types of item formats.

Topic Guide

When the aims of the instrument are specified in advance it is possible to apply an initial structure to the assessment instrument – a *topic guide* format, as indicated in the second row of Table 3.3. Patton (1980), in the context of interviewing, labels this the "interview guide" approach—the guide consists of:

> a set of issues that are to be explored with each respondent before interviewing begins. The issues in the outline need not be taken in any particular order and the actual wording of questions to elicit responses about those issues is not determined in advance. The interview guide simply serves as a basic checklist during the interview to make sure that there is common information that should be obtained from each person interviewed. (p. 198)

Table 3.3 Levels of prespecification in item formats

	Intent to measure	Description of item components		Specific items		
Item format	construct "X"	General	Specific	No score guide	Score guide	Responses
Participant observations	**Before or after**	After	After	After	After	After
Topics guide (a): general	Before	**Before**	After	After	After	After
Topics guide (b): specific	Before	Before	**Before**	After	After	After
Open-ended	Before	Before	Before	**Before**	After	After
Open-ended plus scoring guide	Before	Before	Before	Before	**Before**	After
Fixed-response	Before	Before	Before	Before	Before	**Before**

Two levels of specificity in this format are distinguished. At the more general level, the components, including the definition of the construct, are specified only to a summary level—this is called the *general* topic guide approach. In practice, the full specification of these will happen after observations have been made. At the higher level of specificity, the complete set of components, including the construct definition, is available before administration—hence this is called the *specific* topic guide approach. The distinction between these two levels is a matter of degree—one could have a very vague summary or there could be a more detailed summary that was nevertheless incomplete.

Open-Ended

The next level of prespecification is the *open-ended* format. This includes the common forms of open-ended items, interviews, and essay questions. Here, the items are determined before the administration of the instrument and are administered under standard conditions, in a predetermined order. In the context of interviewing, Patton (1980) has labeled this the "standardized open-ended interview." Like the previous level of item format, there are two discernible levels within this category. At the first level, the response categories are yet to be determined. Most tests that teachers make themselves and use in their classrooms are at this level. At the second level, the categories that the responses will be divided into are predetermined—this is called the *scoring guide* level.

Standardized Fixed-Response

The final level of specificity is the *standardized fixed-response* format typified by multiple choice and other forced-choice items. Here, the student *chooses* a response to the item rather than generating one. As mentioned above, this is probably the most widely used form in published instruments. Any multiple-choice instrument is an example.

The foregoing typology is not merely a way to classify the items in instruments that one might come across in research and practice. Its real strength lies in its nature as a guide to the item generation process. It could be argued that every instrument should go through a set of developmental stages that will approximate the columns in Table 3.3 until the desired level is reached. Instrument development efforts that skip levels will often end up having to make more or less arbitrary decisions about item design components at some point. For example, deciding to create a fixed-response type of item without first investigating the responses that people would make to open-ended prompts will leave no defense against the criticism that the fixed-response format has distorted the measurement.

In the next section, we take up the relationship of this task discussion with the needs of performance assessment, commonly claimed as a feature typical of assessments of twenty-first century skills.

New Tasks for Twenty-First Century Skills

The assessment of twenty-first century skills presents many challenges in terms of the characteristics of the evidence required to draw valid inferences. As pointed out in Chap. 2, this will include performance-based assessments. Pursuing traditional paths of argument in assessment may lead to issues of cost, human scoring, interrater reliability, logistics of managing extensive work products, and so forth.

Performance assessment has been defined in many ways over the years. To connect the ideas of tasks in the prior section to performance assessment in this section, we draw on the following quote from Palm (2008):

> Most definitions offered for performance assessment can be viewed as response-centred or simulation-centred. The response-centred definitions focus on the response format of the assessment, and the simulation-centred definitions focus on the observed student performance, requiring that it is similar to the type of performance of interest. (p. 4)

The typology described above speaks to the response-centered definitions of performance needs for twenty-first century skills—what formats allow for appropriate response types? The degree of match between construct and task design can address simulation-centered definitions—what task appropriately simulates the conditions that will inform us about the underlying construct?

New approaches to technology-mediated content such as “assessment objects” which are online learning objects specifically designed for evidence collection, simulations, virtual worlds, sensors, and other virtual capabilities also expand what we might mean by performance-based opportunities for twenty-first century contexts. Such approaches definitely invite more extensive research on their evidence qualities.

Entities such as the growing “digital” divisions of the major educational publishing houses are beginning to embed online assessment opportunities in their products and are being accepted by school districts as part of the standard curriculum adoption process. All of these initiatives mean that there are many new opportunities for the measurement of complex constructs and for the generation of huge amounts of data, should the planned sharing of data across contexts be implemented.

It is likely that new types of performance are now measurable, given computer-mediated interactions and other technology platforms, performances that may suggest new acceptable routes to defining evidence, without incurring the same substantial barriers as was previously the case for entirely paper-and-pencil performance assessments.

Combining Summative and Formative

One important development is the increased ability, because of improved data handling tools and technology connectivity, to combine formative and summative assessment interpretations to give a more complete picture of student learning. Teachers in the classroom are already working with an enormous amount of assessment data that is often performance-related. If good routes for transmitting information between classroom-based and large-scale settings can be identified, this will be a critical advance in the feasibility of measuring twenty-first century skills in performance-based approaches.

It is not a luxury, but almost a necessity, to begin to combine evidence of practices in defensible ways if the goal is to measure twenty-first century skills. Here, the availability of possibly very dense data may be the key to effective practices, although data density alone does not overcome the issues that are raised in this chapter concerning the need for evidence. However, the potentially available—but currently relatively untapped—evidence from classrooms, along with the vastly increased opportunities for efficient and effective data collection offered by technology, means that much more evidence can be made available for understanding student learning. This assumes, of course, that such data are collected in a way that maintains their status as evidence and that sufficient technology is available in schools and perhaps even in homes.

Wisdom of the Crowd

As mentioned previously, new forms of assessment based on “wisdom of the crowd” and like-minded ideas may also expand what counts as evidence. “Ask the customer” has been a long-standing practice in assessment, as in evaluations, survey design, response processes such as exit interviews, and focus groups. The concept of crowd wisdom extends these to a wider reach and much larger data banks of group data, both for normative comparisons on the fly such as the use of iclickers, cross context ratings, and much better ability to combine and retain "historic" data because of enhanced data density and large capacity data storage/processing.

Task Analysis

With enhanced data density, it is now possible to carry out detailed cognitive task analyses of complex performances (Lesgold 2009). Learning by doing can be assessed, along with such questions as persistence and mastery within a complex task. Larger tasks can provide meaningful opportunities for learning as well as assessment, and it may be possible to assess subject matter knowledge at the same time within the same tasks. If this were the case, then testing need not take as much time away from learning, and feedback cycles could be incorporated so that assessment would lead directly to tailored intervention, making the test into part of the learning process, when appropriate.

An example is FREETEXT: French in Context.[7] A data-driven system, FREETEXT uses natural language processing and adaptive hypermedia for second language acquisition through task-based activities. Natural language processing systems such as this can help find, filter, and format information to be displayed in adaptive hypermedia systems, whether for education or for other purposes. Some projects have effectively brought together data collected from whole-language interactions and used this in adaptive hypermedia. Oberlander (2006) describes how this is possible, particularly in what he describes as formatting or information presentation. Here, "natural language generation systems have allowed quite fine-grained personalisation of information to the language, interests and history of individual users" (Oberlander 2006, p. 20).

Embedded Items

This kind of approach is among those that suggest that it may be possible to capture effectively useful assessment results in substantial tasks. Lesgold explains that "the big change would be that items were discovered within meatier cognitive performances" (Lesgold 2009, p. 20). Of course, this may also require more advanced measurement models of various types. One common example is where several items (not necessarily all the items in a test) are based on the reading of a common stimulus passage. This induces a dependency among those specific items that is not controlled for in standard measurement models. This same effect, sometimes called bundle dependency (because these items form a "bundle"), can also be induced when all of the items relate to a specific larger task in the test, and so on. Describing the actual models to handle is beyond the scope of this chapter; the reader is referred to Rosenbaum (1988), Scalise and Wilson (2006, 2007), and Wilson and Adams (1995).

[7] ftp://ftp.cordis.europa.eu/pub/ist/docs/ka3/eat/FREETEXT.pdf

In developing effective tasks, one suggestion is to look at actual student work produced in projects and assignments, and then analyze which proficiencies have been clearly demonstrated and which need additional assessment. Indicators could take some of the new forms described above, but could also be in traditional formats. Depending on the purpose and context of the assessment, common traditional formats that might continue to offer great value if properly incorporated include multiple-choice, short answer, and constructed-response essays. It is likely to be found that tasks could contain a variety of formats, mixing innovative and more conventional approaches. Libraries of such tasks might accumulate, both created by teachers and instructors and made available to them as a shared resource, and others could be retained for larger-scale settings. An interesting example of the ways that technology can support the use of embedded tasks is found in the Package Tracer software used in Cisco's Networking Academy.

The Packet Tracer (PT) used in the Cisco Networking Academy program is a comprehensive simulation and assessment environment for teaching networking concepts (Frezzo et al. 2009, 2010). An important aspect of the PT is the integration between curriculum and assessment, which allows the collection of evidence of the students' learning through the instructional objects that are developed in the PT, in other words, it is not necessary to create assessment tools that are distinctively separate from the typical objects used during instruction in order to inform student assessments (Frezzo et al. 2010).

The PT software is intended to develop through instruction the competencies that characterize network engineers. With this purpose in mind, the software presents students with simulations related to their instructional goals (Frezzo et al. 2009, 2010).

The simulations in the PT are presented through a navigable interface that supports the presentation of information and scenarios and allows the interaction of students with those scenarios. Figure 3.7 presents a sample screenshot of the interface of the PT (Frezzo et al. 2010).

The PT software allows instructors to develop a variety of instructional tasks within this environment including activities that illustrate specific concepts, activities that promote the practice of procedural skills, open-ended tasks that allow for a variety of potential solutions, and finally troubleshooting scenarios that require students to identify problems and develop solutions (Frezzo et al. 2010).

Seamless integration between the learning tasks and assessment is achieved by the association of (a) an instructional task or simulation with (b) an "answer network" (i.e., a prototype or exemplar of what would constitute a functional setup) and a "grading tree" that indicates how different aspects of the "answer network" or exemplar must be valued (Frezzo et al. 2010). In terms of the concepts discussed in this report, the assessment is achieved by linking the instructional tasks with an exemplar of the expected learning performance and an outcome space that informs how the different possible responses should be valued. The key aspect of this kind of assessment is that, by providing this "answer network" (proficiency exemplar) and "grading tree" (an outcome space or scoring guide), it is possible for instructors to create automatically scored assessments based on the same kinds of tasks that they would use for instruction (Frezzo et al. 2010).

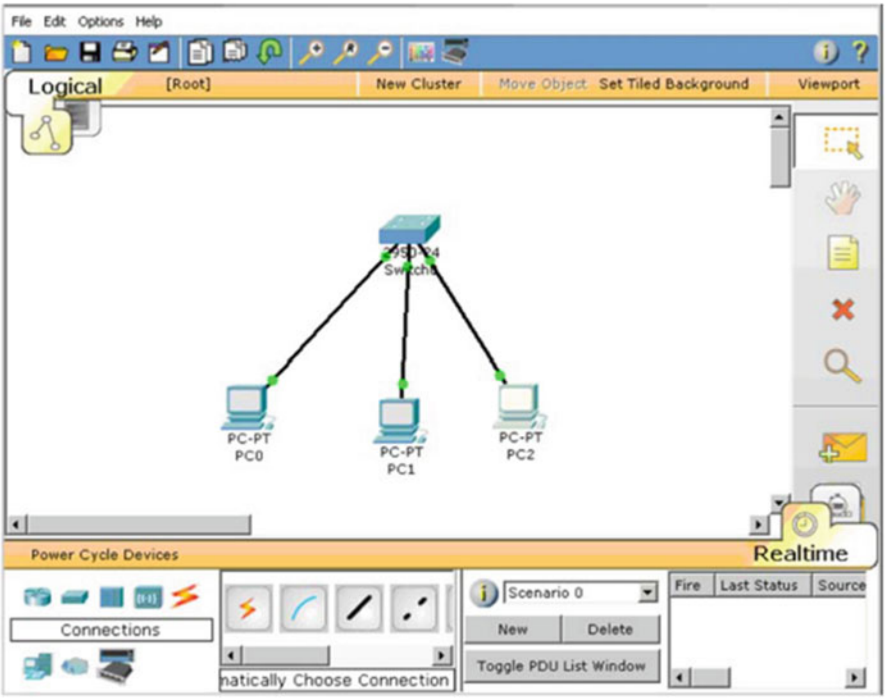
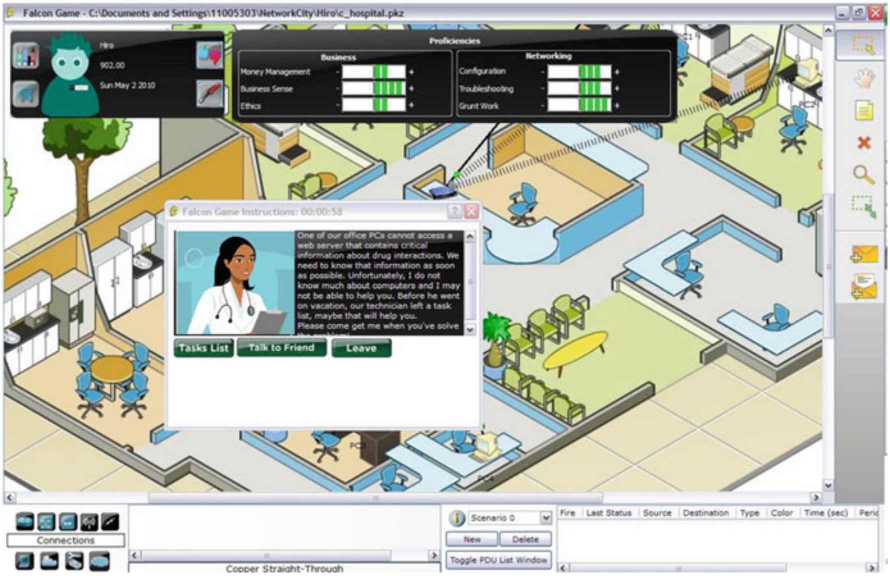


Fig. 3.7 Designing the tasks—screenshot of Packet Tracer used in Cisco Networking Academies (Frezzo et al. 2010)

The PT constitutes an interesting example of the high levels of sophistication that can be achieved in the creation of simulation environments for instructional tasks as well as the possibilities that these kinds of environments offer for the integration of instruction and assessment. Moreover, the possibilities of these kinds of simulation environments can be expanded further by the development of new interfaces to present more challenging scenarios while at the same time making them more intuitive for the students.

An illustration of these new possibilities is given by recent work on the Cisco Networking Academy to integrate the PT with a game-like interface similar to the concepts used in online social games (Behrens et al. 2007). (Figure 3.8 presents a screenshot of this kind of interface.) The use of this kind of environment opens possibilities for developing assessments based on scenarios using different social interactions such as dealing with clients or presenting business proposals. Additionally, this new interface creates opportunities to assess not only proficiency in specific content domains (in this case, proficiency in networking concepts) but also additional competencies that might be involved in more authentic tasks, such as social skills (Behrens et al. 2007).



One of our office PCs cannot access a web server that contains critical information about drug interactions. We need to know that information as soon as possible. Unfortunately, I do not know much about computers and I may not be able to help you. Before he went on vacation, our technician left a task list, maybe that will help you.
Please come get me when you've solve

Fig. 3.8 Screenshot of an example of new interfaces being developed for Cisco networking academies

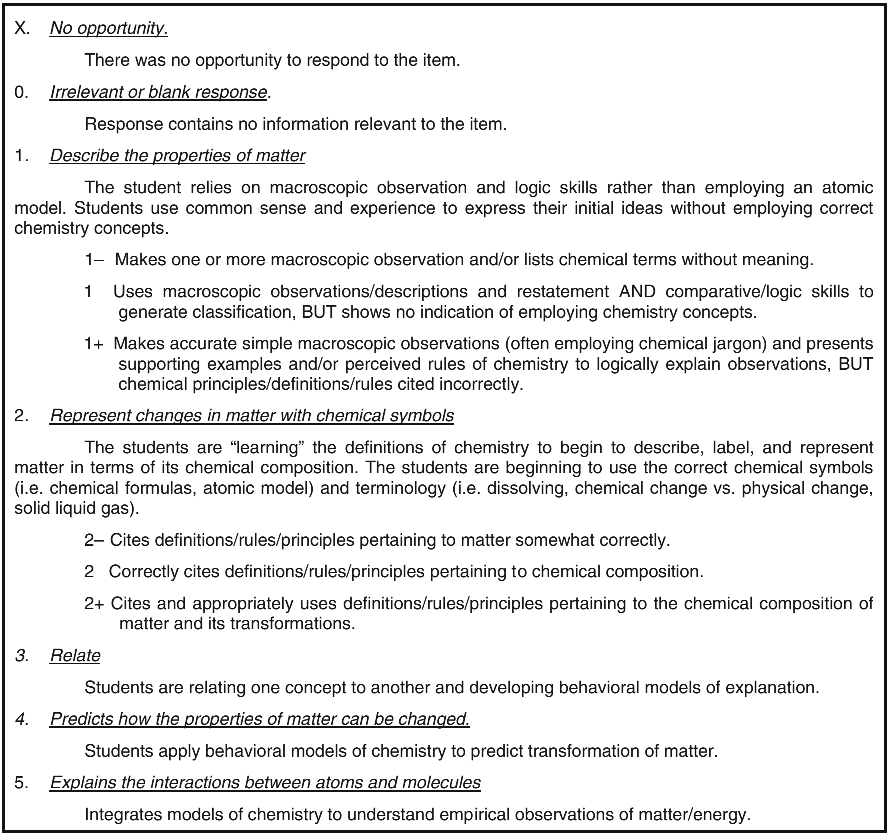
Valuing the Responses[8]

In order to analyze the responses and products collected through the tasks, it is necessary to determine explicitly the qualitatively distinct categories into which student performances can be classified. The definition of distinct categories is commonly operationalized in practice in the form of scoring guides, which allow teachers and raters to organize student responses to assessment tasks.

In much of his writing, Marton (1981, 1983, 1986, 1988, Marton et al. 1984) describes the development of a set of outcome categories as a process of "discovering" the qualitatively different ways in which students respond to a task. In this chapter, we follow the lead of Masters and Wilson (1997), and the term outcome space is adopted and applied in a broader sense to any set of qualitatively described categories for recording and/or judging how respondents have responded to items.

Inherent in the idea of categorization is the understanding that the categories that define the outcome space are qualitatively distinct; in reality, all measures are based, at some point, on such qualitative distinctions. Rasch (1977, p. 68) pointed out that this principle goes far beyond measurement in the social sciences: "That science should require observations to be measurable quantities is a mistake of course; even in physics, observations may be qualitative—as in the last analysis

[8] The following section has been adapted from Wilson 2005.



X. *No opportunity.*

There was no opportunity to respond to the item.

0. *Irrelevant or blank response*.

Response contains no information relevant to the item.

1. *Describe the properties of matter*

The student relies on macroscopic observation and logic skills rather than employing an atomic model. Students use common sense and experience to express their initial ideas without employing correct chemistry concepts.

1– Makes one or more macroscopic observation and/or lists chemical terms without meaning.

1 Uses macroscopic observations/descriptions and restatement AND comparative/logic skills to generate classification, BUT shows no indication of employing chemistry concepts.

1+ Makes accurate simple macroscopic observations (often employing chemical jargon) and presents supporting examples and/or perceived rules of chemistry to logically explain observations, BUT chemical principles/definitions/rules cited incorrectly.

2. *Represent changes in matter with chemical symbols*

The students are “learning” the definitions of chemistry to begin to describe, label, and represent matter in terms of its chemical composition. The students are beginning to use the correct chemical symbols (i.e. chemical formulas, atomic model) and terminology (i.e. dissolving, chemical change vs. physical change, solid liquid gas).

2– Cites definitions/rules/principles pertaining to matter somewhat correctly.

2 Correctly cites definitions/rules/principles pertaining to chemical composition.

2+ Cites and appropriately uses definitions/rules/principles pertaining to the chemical composition of matter and its transformations.

3. *Relate*

Students are relating one concept to another and developing behavioral models of explanation.

4. *Predicts how the properties of matter can be changed.*

Students apply behavioral models of chemistry to predict transformation of matter.

5. *Explains the interactions between atoms and molecules*

Integrates models of chemistry to understand empirical observations of matter/energy.

Fig. 3.9 Outcome space as a scoring guide from the Living by Chemistry Project

they always are.” Dahlgren (1984) describes an outcome space as a “kind of analytic map”:

> It is an empirical concept, which is not the product of logical or deductive analysis, but instead results from intensive examination of empirical data. Equally important, the outcome space is content-specific: the set of descriptive categories arrived at has not been determined a priori, but depends on the specific content of the task. (p. 26)

The characteristics of an *outcome space* are that the categories are well-defined, finite and exhaustive, ordered, context-specific, and research-based.

An example of the use of scoring guides as a representation of the different response categories that comprise the outcome space of a task can be seen in Fig. 3.9. In this case, the construct is “Matter” and is designed to represent levels of student understanding about the role of matter in Chemistry curricula from late high school through early college levels. It has been designed as part of the Living By Chemistry (LBC) project (Claesgens et al. 2009).

Research-Based Categories

The construction of an outcome space should be part of the process of developing an item and, hence, should be informed by research aimed at establishing the construct to be measured, and identifying and understanding the variety of responses students give to that task. In the domain of measuring achievement, a National Research Council (2001) committee has concluded:

> A model of cognition and learning should serve as the cornerstone of the assessment design process. This model should be based on the best available understanding of how students represent knowledge and develop competence in the domain… This model may be fine-grained and very elaborate or more coarsely grained, depending on the purpose of the assessment, but it should always be based on empirical studies of learners in a domain. Ideally, the model will also provide a developmental perspective, showing typical ways in which learners progress toward competence. (pp. 2–5)

Thus, in the achievement context, a research-based model of cognition and learning should be the foundation for the definition of the construct, and hence also for the design of the outcome space and the development of items.

Context-Specific Categories

In the measurement of a construct, the outcome space must always be specific to that construct and to the contexts in which it is to be used.

Finite and Exhaustive Categories

The responses that the measurer obtains to an open-ended item will generally be a sample from a very large population of possible responses. Consider a single essay prompt—something like the classic "What did you do over the summer vacation?" Suppose that there is a restriction to the length of the essay of, say, five pages. Think of how many possible different essays could be written in response to that prompt. Multiply this by number of different possible prompts, and then again by all the different possible sorts of administrative conditions, resulting in an even bigger number. The role of the outcome space is to bring order and sense to this extremely large set of potential responses. One prime characteristic is that the outcome space should consist of only a finite number of categories and, to be fully useful, must also be exhaustive, that there must be a category for every possible response.

Ordered Categories

Additionally, for an outcome space to be informative in defining a construct that is to be mapped, the categories must be capable of being ordered in some way. Some categories must represent lower levels on the construct and some must represent higher ones. This ordering needs to be supported by both the theory behind the construct—the theory behind the outcome space should be the same as that behind the construct itself—and by empirical evidence. Empirical evidence can be used to support the ordering of an outcome space and is an essential part of both pilot and field investigations of an instrument. The ordering of the categories does not need to be complete. An ordered partition (in which several categories can have the same rank in the ordering) can still be used to provide useful information (Wilson and Adams 1995).

The development of an outcome space that meets the four aforementioned criteria allows the performance criteria for the assessments to be clear and explicit—not only to teachers but also to students and parents, administrators, or other "consumers" of assessment results. The use of clear and explicit scoring criteria is an important element that can lend credibility to the inferences based on the assessment process by making transparent the relation between the tasks, the responses, and the construct.

Valuing the Responses—Example: The Using Evidence Framework

The relevance of a cognitive model as the starting point for an assessment is related to its role as the base for interpreting and evaluating students' products and responses. By modeling the individual components and processes involved in scientific reasoning, the Using Evidence framework (UE; 2008, 2010a, 2010b, introduced previously as an example in "Defining the constructs") supports the assessment and analysis of multiple facets of this process (Brown et al. 2010a).

For example, in the UE model, the "rules" component can be assessed in terms of the "accuracy" of rules that the students are using when thinking about evidence. The assessment of the accuracy of the rules as a measure of quality can then be instantiated in terms of a wide variety of formats, ranging from simple correct/incorrect dichotomous items to a scoring guide that captures the level of sophistication of the rules used by students (Brown et al. 2010a).

An example provided by Brown et al. (2010a) serves to illustrate how a scoring guide would capture those relative levels. Consider the three following statements:

a) "something that is dense will sink"
b) "something that is heavy will sink"
c) "something with holes will sink"

Although none of these statements are fully accurate, it is still possible to associate them with three ordered levels of proficiency, where rule (a) seems to indicate a more nuanced understanding than (b), and (b) seems to indicate a higher level of proficiency than (c).

The Conceptual Sophistication construct developed in the UE framework attempts to capture quality and complexity of student responses ranging from

Table 3.4 Conceptual sophistication outcome space (Brown et al. 2010b)

Response category	Description	Example responses
Multicombined	Applying one concept derived from combined concepts	"It will sink if the *relative density* is large"
Multirelational	Relating more than one combined concept	"It will sink if the *density of the object is greater than the density of the medium*"
Combined	Applying one concept derived from primary concepts	"It will sink if the *density* is large"
Relational	Relating more than one primary concept	"It will sink if the *mass is greater than the volume*" "It will sink if the *buoyant force is less than the gravitational force*"
Singular	Applying one primary concept	"It will sink if the *mass* is large" "It will sink if the *volume* is small" "It will sink if the *buoyant force* is small"
Productive misconception	Applying one or more non-normative concepts that provide a good foundation for further instruction	"It will sink if it's *heavy*" "It will sink if it's *big*" "It will sink if it's not *hollow*"
Unproductive misconception	Applying one or more non-normative concepts that provide a poor foundation for further instruction	"It will sink if it's not *flat*" "It will sink if it has *holes*"

Table 3.5 Function of a statement and its relationship to surrounding statements (Brown et al., 2010a)

Statement	Function in argument	Surrounding statements	Function in argument
"this block is heavy…"	Premise	"…therefore it will sink"	Claim
"this block is heavy…"	Claim	"…because it sank"	Premise
"this block is heavy…"	Part of datum	"…and it sank"	Part of datum

misconception at the lower level up to the coordination of a multiplicity of ideas that support normative scientific conceptions (Brown et al. 2010b). Table 3.4 presents a summary of the different levels of the Conceptual Sophistication construct and illustrates the articulation of the levels of the cognitive progression (in the response category column and its description) and the student responses.

Another application of the UE framework that is interesting to note is how the model can be linked to and used to organize specific aspects of the evaluation of student responses. The model can give a structure to consider the location and purpose of the statement within the context of an entire argument presented by the students by capturing, for example, that the function of a statement can vary depending on its relationship to surrounding statements, providing valuable information about the process of reasoning employed by the students (Brown et al. 2010a). A simple example of this kind of distinction is presented in Table 3.5.

Delivering the Tasks and Gathering the Responses

An important aspect of operationalizing an assessment is the medium of delivery. The decision to rely on computers for task delivery and response gathering influences many design questions, and therefore, this decision should take place very early on. For example, one of the many opportunities that computer delivery opens up is the possibility of automated scoring (Williamson et al. 2006) of constructed test responses because the response—an essay, speech sample, or other work product—is available digitally as a by-product of the testing process. However, as is the case with traditional forms of assessments, as scholars and researchers (Almond et al. 2002; Bennett and Bejar 1998) have noted, in order to take full advantage of the benefits of automated scoring, all other aspects of the assessment should be designed in concert.

Although there appears to be little doubt that computer test delivery will be the norm eventually, some challenges remain to be solved. Perhaps the most sobering lesson learnt from the use of the computer as a delivery medium for large-scale testing has been that there is a capacity problem. The capacity or examinee access problem refers to the lack of sufficient number of testing stations to test all students at once (Wainer and Dorans 2000, p. 272). In contrast, large-scale paper-and-pencil testing of large student populations is routinely carried out across the world, even in very poor countries. If the assessment calls for a large number of students to be tested at once, the paper-and-pencil medium still remains the likely choice. Eventually, the increasing availability of technology in the form of a multiplicity of portable devices as well as the decreasing costs of computers should solve the issues of capacity. One possibility is to use the student's own computer as a testing terminal, although some problems would need to be taken into account. For one thing, a wide variety of computers exist, which may preclude sufficiently standardized testing conditions. In addition, for tests where security is necessary, the use of student computers could present a security risk. An additional consideration is connectivity (Drasgow et al. 2006, p. 484). Even if the computers or alternative devices are available, they need to be supplied with information to carry out the testing process. In turn, local devices serving as a testing station need to forward information to a central location. Unless connectivity between the local computers and the central location is extensive and reliable, the testing process can be disrupted, which can be especially detrimental in a context of high-stakes assessment.

As an alternative to the capacity problem or as an additional solution, the testing can be distributed over many testing occasions. For example, the TOEFL (Test of English as a Foreign Language) is administered globally every week. In this case, taking the exam involves a process not unlike making reservations on an airline: it is necessary to make an appointment or reservation to take the exam on a specific administration date (Drasgow et al. 2006, p. 481). Distributing the assessment over multiple administration days goes a long way toward solving the problem of limited capacity, but in reality, it just ameliorates the problem since some dates, as is the case with flight reservations, are more popular than others. When a preferred date is not available, the students need to be tested on an alternative date. Of course, it also means that the test design must deal with the fact that the content of the test is constantly being revealed to successive waves of students.

One of the major advantages of computer test delivery is that the assessment can possibly be designed to be adaptive. Computerized adaptive testing (CAT) was the earliest attempt to design an assessment that went beyond merely displaying the items on the computer screen; the early research on this idea was carried by Lord (1971). Since then, the approach has been used operationally in several testing programs (Drasgow et al. 2006, p. 490), and research continues unabated (Van der Linden and Glas 2007; Weiss 2007).

Adaptive testing has raised its own set of challenges; one that has received attention from researchers throughout the world is so-called exposure control, which refers to the fact that items in an item pool could be presented so frequently that the risk of "exposing" the item becomes unacceptable. In fact, items with particularly appropriate qualities for the task at hand tend to be selected more frequently by any automated item selection procedure—good items tend to get used up faster. Overexposed items effectively become released items, so that subsequent test takers could have an advantage over earlier test takers. The interpretation of scores, as a result, can be eroded over time. Multiple solutions to this problem have been offered, and an overview can be found in Drasgow et al. (2006, p. 489). One solution is to prevent items from being overexposed in the first place by designing the item selection algorithm in such a way as to distribute the exposure to all items equally without reducing the precision of the resulting ability estimates. An alternative solution is to effectively create so many items that the chance of exposure of any of them is diminished considerably (Bejar et al. 2003).

An additional approach to addressing exposure in CAT is to create tasks of sufficient complexity that they can be exposed even to the degree of complete transparency of item banks, without increasing the likelihood of a correct response in the absence of sufficient construct proficiency (Scalise 2004). While this is somewhat of a look-ahead, given the methodological issues with complex tasks described in this chapter, twenty-first century skills and tasks may be ideally suited for this "transparent" exposure approach, given sufficient research and validation over time.

Despite the challenges, the potential advantages of computer test delivery are numerous and very appealing. Among the advantages are the possibility of increased convenience to the test taker and the possibility of much faster turnaround of test results. The delivery of test by computer creates opportunities to enhance what is being measured, although taking advantage of that opportunity is not simply a matter of delivery; the assessment as a whole needs to be designed to take advantage of the possibilities offered by computer delivery.

For example, item formats that go beyond the multiple-choice format could offer more valid assessments, provided that irrelevant variability is not introduced in the process. As noted earlier, constructed responses can be captured digitally as a by-product of computer test delivery, and, therefore, their scoring can be greatly facilitated, whether scored by judges or by automated means. Online scoring networks (Mislevy et al. 2008) have been developed that can score, in relatively short order, constructed responses across time zones and by judges with different backgrounds. Automated scoring of written responses is a reality (Drasgow et al. 2006, p. 493), and the automated scoring of speech is advancing rapidly (Zechner et al. 2009). Similarly, the automated scoring of some professional assessments (Braun et al. 2006; Margolis and Clauser 2006) has been used for some time.

Of special interest for assessment of twenty-first century skills is the assessment of what might be called collaborative skills. The need for such skills arises from the demands in the work place for collaboration. Cross-national alliances between corporations, for example, are seen as critical in an increasingly global economy (Kanter 1994). An armchair job analysis of such requirements suggests the need for communication skills that go beyond the purely linguistic skills measured by admissions-oriented assessments like the TOEFL. Instead, the communication skills that need to be developed and assessed are far more subtle. For example, linguists have proposed the term “speech acts” to describe the recurring communicative exchanges that take place in specific settings (Searle 1969) and involve at least two protagonists. The content of what is said in those exchanges is certainly important but so is *how* it is said, which is function of the role of the protagonists, the background information they share in common, and so forth. The “how” includes attributes of the speech, for example, tone, but also “body language” and more importantly facial expression. An approach to designing assessments at this level of complexity could rely on the extensive work, by Weekley and Ployhart (2006), on situational judgment tests (SJTs), although much more is needed. Interestingly, since collaborative exchanges are increasingly computer-mediated, the assessment of collaborative skills through computer test delivery can be quite natural. One form that collaboration can take is the online or virtual meeting; a form of assessment that simulates an online meeting would be a reasonable approach. For example, in an SJT-based approach, the item could start with a snippet of an online exchange as the stimulus for the student, and the test taker would then need to offer some judgment about it, while a more advanced approach would have the students contribute to the exchange at selected points. What the student says, how he says it, and his/her facial expression and body language would all be part of the “response.” Progress along these lines is already appearing (Graesser et al. 2007).

Modeling the Responses

A key to understanding the proficiency status or states of knowledge of students is recognizing the intricacies of any testing data collected in the process. In subsequent sections of this chapter, we advocate the reporting of results to users at all levels, from students and teachers up through school administrators and beyond.

The ability to report such results, however, depends upon the type of assessment given and its scope. For assessments administered to all students in a region (such as end-of-grade tests given in the USA), such reports are possible. For other tests that use intricate sampling designs, such as the US National Assessment of Educational Progress (NAEP), many levels exist for which reports are not possible. For instance, in NAEP, student and school reports are purposely omitted due to a lack of adequate samples for the estimates at each level (samples in terms of content in the former case and numbers of students in the latter).

The key to understanding what is possible is a thorough grasp of the statistical issues associated with the sampling and measurement design of the assessment. Consequently, statistical methods and models that incorporate such information are valuable tools only to the extent that they comply with the demands of the context and the patterns of the data. In cases where group results are the aim, techniques such as weighted analyses and/or multilevel models (or hierarchical linear models) that allow for the dependencies of clustered data to be represented should be used in an analysis. Of course, these weights can be inconsistent with the usual concept of fairness in testing, where each individual is judged only by performance. Regardless of the form chosen, it is important to plan for the use of these models at all stages of the test development process, so as to guide decisions about the type and scope of the sampling and measurement design.

To demonstrate what we mean by the modeling of responses, we describe an example based on the papers of Henson and Templin (2008) and Templin and Henson (2008)—this example will also be used below to describe an example of a report to users and the remedial actions to which they might lead. The authors used a Diagnostic Classification Model (or DCM; see Rupp and Templin (2008)) to analyze a low-stakes formative test of Algebra developed for an impoverished urban school district in a southeastern American state. DCMs are psychometric models that attempt to provide multidimensional feedback on the current knowledge state of an individual. DCMs treat each trait as a dichotomy—either students have demonstrated mastery of a particular content area or they have not. We highlight DCMs not to suggest them as psychometric models but to show how psychometrics can lead to actionable result reporting.

The data underlying this example come from a 25-item benchmark test of basic 3rd grade science skills (Ackerman et al. 2006), used to diagnose students' mastery of five basic science skills. For instance, a student in the report might have a high probability of mastering the skills associated with "Systems" (.97), "Classification" (.94), and "Prediction" (.97), but she will most likely not master "Measurement" (.07). For the "Observation" skill, she has a probability of .45 of being a master, making her diagnosis on that skill uncertain.

In Henson and Templin (2008), the authors used a standard setting procedure to create classification rules for evaluating mastery status of students on five skills associated with Algebra. The formative test was built to mimic the five skills most represented in the state end-of-grade examination, with the intent being to provide each student and teacher with a profile of the skills needed to succeed in Algebra according to the standards for the State.

Templin and Henson (2008) reported the process of linking student mastery profiles with the end-of-grade test, shown to demonstrate how such reports can lead to direct actions. Students took the formative assessment in the middle of the academic year and took the end-of-grade assessment at the end of the year. The students' mastery profiles from the formative tests were then linked with their performance on the end-of-grade assessment.

For the State, the goal was to make each student reach the State standard for proficiency in Algebra, which represented a score of approximately 33 out of the 50 item end-of-grade assessment. By linking the formative mastery profiles with the end-of-grade data, Templin and Henson were able to quantify the impact of acquiring mastery of each of the attributes in terms of increase in test score on the end-of-grade assessment. Figure 3.10 shows a network graph of all 32 possible mastery statuses (combinations of mastery or nonmastery for all five Algebra skills). Each master status (shown as the nodes of the graph) is linked to the status that has the highest increase in end-of-grade test score for the status where one additional attribute is mastered. For example, the node on the far right of the graph represents the mastery status where only the fifth skill was mastered. This node is connected to the status where the fifth and second skill have been mastered—indicating that students who have only mastered the fifth skill should study the second skill to maximize their increase in end-of-grade test score.

Figure 3.11 is a rerepresentation of the network graph shown in Fig. 3.10, this time superimposed on the scale of the end-of-grade test score. Each of the example students shown in Fig. 6 is given a "pathway to proficiency"—a remediation strategy that will be the fastest path to becoming proficient, in terms of the State criterion. For instance, Student A, who has not mastered any skills, should work to learn skill two, then skill one. Although such pathways must be tailored to fit each scenario with respect to timing in the curriculum, nature of cognition, and outcome measure (i.e., proficiency need not be defined as a cutscore on an end-of-grade test), such types of reports can lead to actions that will help remediate students and provide more utility for test results.

Modeling the Responses—Example: The Using Evidence Framework

After a construct has been defined, items have been developed, and answers to those have been collected and scored, the next step is to apply a measurement model that will allow us to make an inference regarding the level of proficiency of a student or respondent in general. The Using Evidence framework example (Brown et al. 2008, 2010a, 2010b), previously introduced, will also serve as an example illustrating different issues that should be considered when applying a measurement model, such as the original hypothesis regarding the structure of the construct (whether it is continuous or categorical, for example), the nature of the responses being modeled, and the unidimensional or multidimensionality of the construct, among others. In the case of the UE framework, the construct was defined as multidimensional in

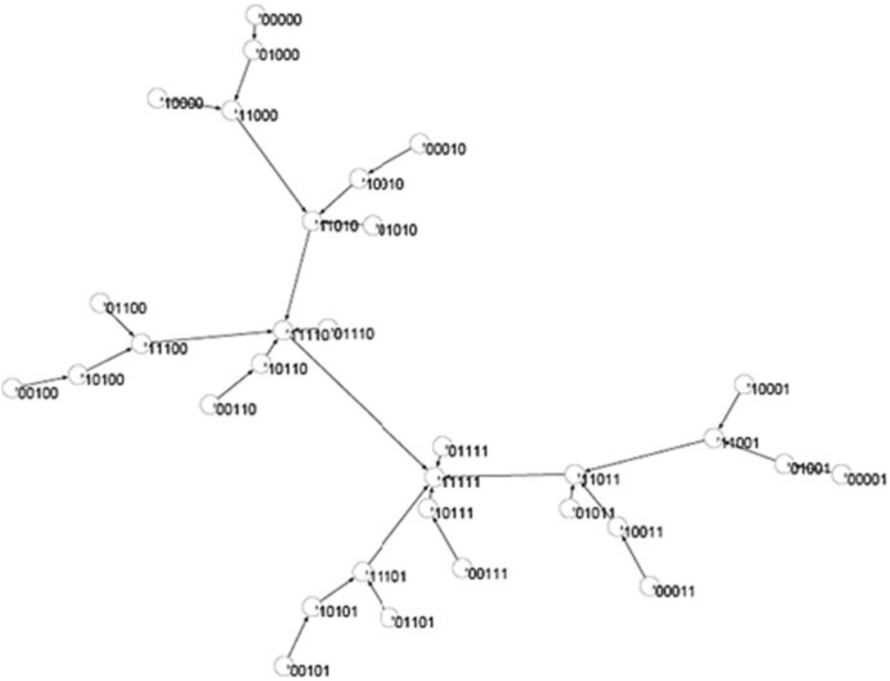


Fig. 3.10 Proficiency road map of binary attributes

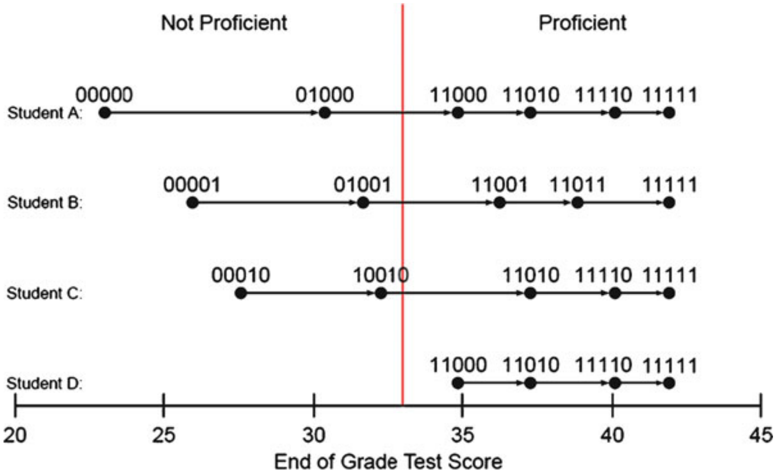


Fig. 3.11 Fast path to proficiency

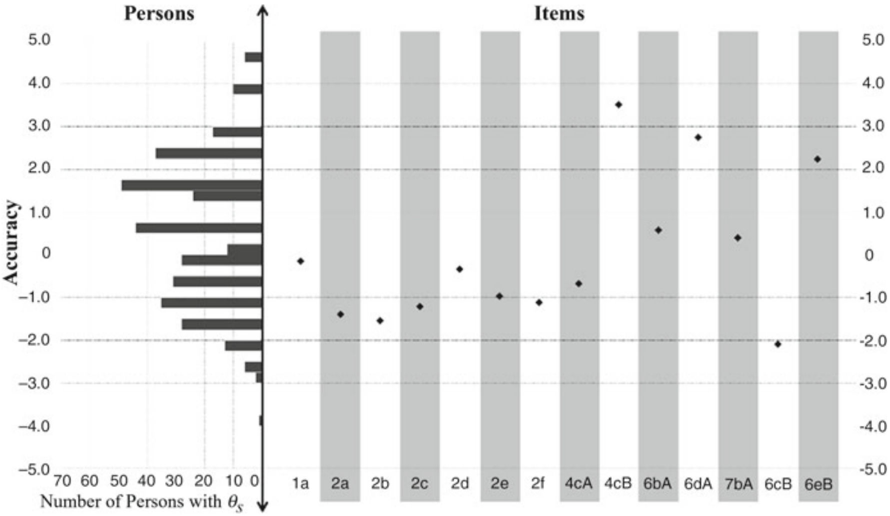


Fig. 3.12 Wright map for dichotomous items in the accuracy construct (Brown et al. 2010b)

nature, and the test used to collect student responses contained both dichotomous and polytomous items. The simplest form of response were the dichotomous items used in the "Accuracy" dimension of the UE framework. In order to model these responses, a Rasch (1960/1980) simple logistic model (also known as the 1-parameter logistic model) was used. For this model, the probability that a student answers a particular item correctly depends on two elements, the difficulty of that item and the proficiency of the student. The probability of a correct response is then modeled as a function of the difference between these two elements:

$$\text{Probability of a correct response of student } j \text{ on item } i = \\ \text{f}(\text{Proficiency of student } j - \text{ Difficulty of item } i)$$

When the difference between student proficiency and item difficulty is 0 (i.e., they are equal), the student will have a probability of .5 of answering the item correctly. If the difference is positive (when student proficiency is greater than the item difficulty), the student will have a higher probability of getting the item correct, and when the difference is negative (when the item difficulty is greater), the student will have a lower probability of answering the item correctly. Using this model, we can represent each item by its difficulty and each student by its proficiency on a single scale, which allows us to use a powerful yet simple graphical tool that can be used to represent the parameters, the Wright map (named after Ben Wright). In this representation, item difficulties and person proficiencies are displayed on the same scale, facilitating the comparison between items and the analysis of their overall relation to the proficiency of the respondents. An example of a Wright map for the accuracy dimension (Brown et al., 2010b) containing only dichotomous items is presented in Fig. 3.12

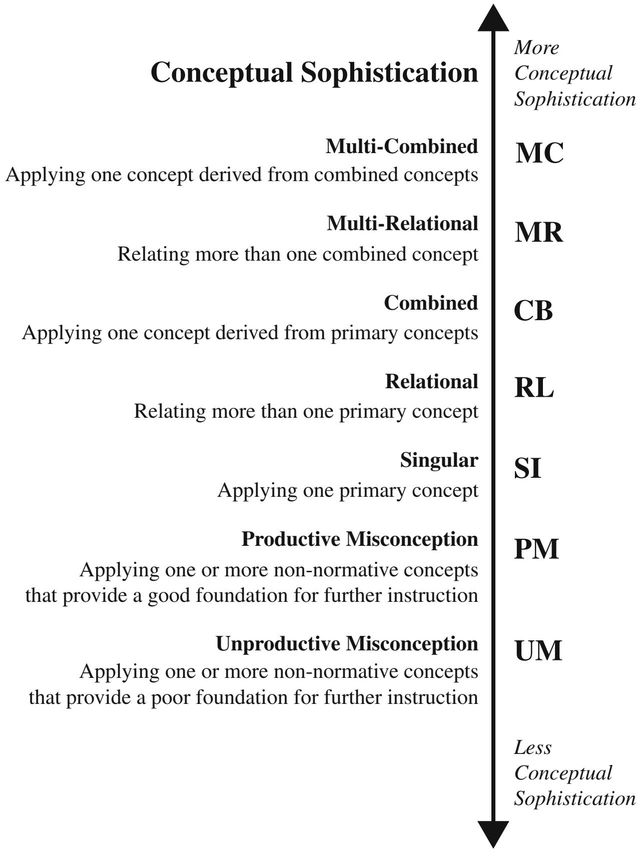

Conceptual Sophistication
More Conceptual Sophistication
Multi-Combined
Applying one concept derived from combined concepts
MC
Multi-Relational
Relating more than one combined concept
MR
Combined
Applying one concept derived from primary concepts
CB
Relational
Relating more than one primary concept
RL
Singular
Applying one primary concept
SI
Productive Misconception
Applying one or more non-normative concepts that provide a good foundation for further instruction
PM
Unproductive Misconception
Applying one or more non-normative concepts that provide a poor foundation for further instruction
UM
Less Conceptual Sophistication

Fig. 3.13 Items in the Conceptual Sophistication construct (Brown et al. 2010b)

On the left side of Fig. 3.12 is a rotated histogram indicating the distribution of the proficiency estimates for the students, and on the right side the difficulty estimates for 14 different items (each being represented by a single point in the scale). Going back to the interpretation of these parameters, one can interpret that students located at the same level as an item will have a probability of .5 of responding it correctly. When an item is located above a student, that means that the student has a probability lower than .5 of answering correctly, and vice versa if the item is located below the student. In other words, it is possible to quickly identify difficult items, namely items that are above most of the students (item 4cB for example), as well as to locate easier items, corresponding to items that are below most of the students (item 6cB for example).

The use of these models allows connection of the model results with the original definition of the construct. As an example of this connection, in the UE framework, we can revisit the "Conceptual Sophistication" construct (Brown et al. 2010b), which defined seven different levels in which a student response could be classified. The Conceptual Sophistication construct is presented in Fig. 3.13

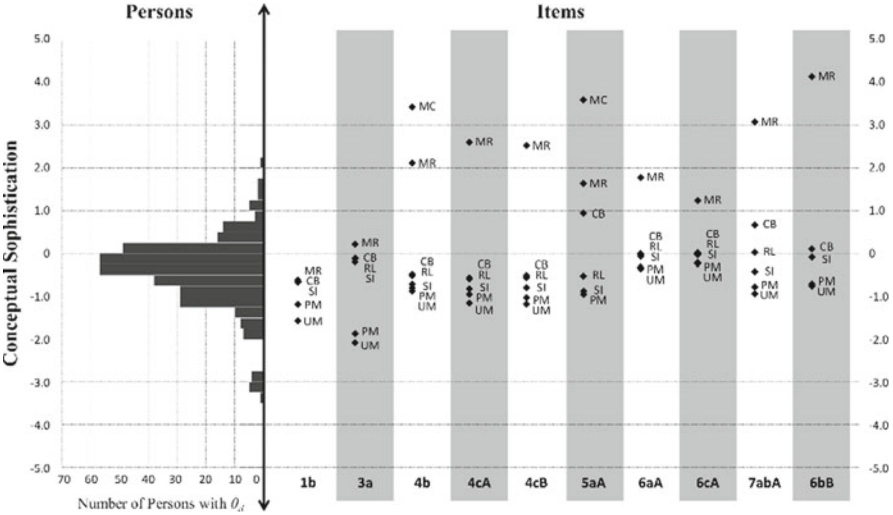


Fig. 3.14 Wright map for polytomous items in the Conceptual Sophistication construct (Brown et al. 2010b)

Under this construct, the answers to the items can be categorized in any of these seven levels, hence, the items will, in general, be polytomous. This kind of item can be analyzed with Masters' (1982) partial credit model (PCM), a polytomous extension of the Rasch model. Within the PCM, a polytomous item, an item with *n* categories, is modeled in terms of *n* − *1* comparisons between the categories. When we represent these graphically, we use the Thurstonian thresholds which indicate the successive points on in the proficiency scale where a response at a level *k* or above becomes as likely as a response at *k* − *1* or below. Figure 3.14 presents the results of a PCM analysis using a modified Wright map which connects the empirical locations of the different Thurstonian thresholds for each Conceptual Sophistication item with the distribution of person estimates. Here we can see that there is some consistency, and also some variation, in the levels of the thresholds for different items. For example, for most items, only students above about 1 logit on the scale are likely to give a response at the multirelational level, and there are relatively few students above that point. However, items *1b* and *3a* seem to generate such a response at a lower level—this may be of great importance for someone designing a set of formative assessments of this construct. Note that not all levels are shown for each item—this occurs when some levels are not found among the responses for that item.

Validity Evidence

The Standards for Educational and Psychological Testing (American Psychological Association, American Educational Association and National Council for Measurement in Education 1985) describe different sources of evidence of validity that need to be integrated to form a coherent validity argument. These include evidence based on test content, response process, internal test structure, relations to other variables, and testing consequences. In earlier sections of this chapter, in particular, those on the construct, the tasks, the outcome space, and the modeling of student responses, we have already discussed aspects of evidence based on test content, response process, and internal test structure. In the two sections below, we discuss aspects of evidence concerning relations to other variables and testing consequences (in the sense of reports to users).

Relations to Other Variables

Some though not all of the twenty-first century skills have attributes that are less cognitive than traditional academic competencies. One of the problems that has bedeviled attempts to produce assessments for such skills (e.g., leadership, collaboration, assertiveness, "interpersonal skills") is that it appears to be difficult to design assessments that are resistant to attempts to "game" the system (Kyllonen et al. 2005). Where such assessments are used in "low-stakes" settings, this is not likely to be much of an issue. But if the assessments of twenty-first century skills are "low-stakes," then their impact on educational systems may be limited. There has been some debate as to whether the possibility that the effects of assessment outcomes can change when they are used in high-stakes settings should be viewed as an aspect of the validity of the assessment. Popham (1997) suggested that while such features of the implementation of assessments were important, it was not appropriate to extend the meaning of the term validity to cover this aspect of assessment use. Messick (1989) proposed that the social consequences of test use should be regarded as an aspect of validity under certain, carefully prescribed conditions.

> As has been stressed several times already, it is not that adverse social consequences of test use render that use invalid but rather that adverse social consequences should not be attributable to any source of test invalidity such as construct-irrelevant variance. If adverse social consequences are empirically traceable to sources of test invalidity, then the validity of the test use is jeopardized. If the social consequences cannot be so traced—or if the validation process can discount sources of test invalidity as the likely determinants, or at least render them less plausible—then the validity of the test use is not overturned. Adverse social consequences associated with valid test interpretation and use may implicate the attributes validly assessed as they function under the existing social conditions of the applied setting, but they are not in themselves indicative of invalidity (Messick 1989 p. 88–89).

In other words, you cannot blame the messenger if the message (accurately) delivers a message that has negative consequences.

Reporting to Users

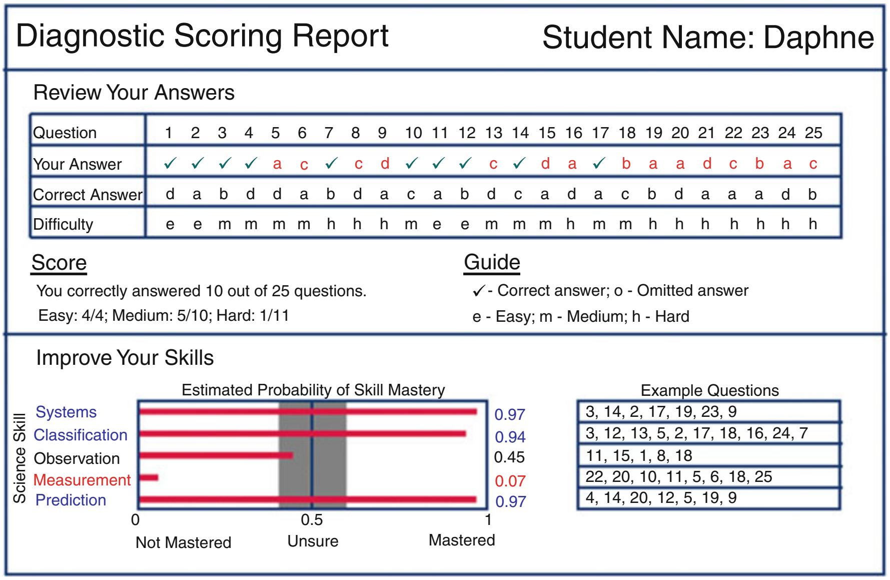
A key aspect of assessment validity is the form that results take when being reported to users. Depending on the scale of the testing program, results currently are reported either immediately following a test (for a computerized test) or are delivered after a (fairly short) period of time. For smaller-scale tests, results may be available nearly immediately, depending on the scoring guide for the test itself.

It is from the results of an assessment that interpretations and decisions are made which influence the knowledge acquisition of test takers and the directions of teaching. Therefore, it is critically important for the success of any assessment system to include features that allow end users to evaluate and affect progress beyond the level of reporting that has been traditional in testing. To that end, we suggest that results of assessments must include several key features. First and foremost, results must be *actionable*. Specifically, they must be presented in a manner that is easily interpretable by end users and can directly lead to actions that will improve targeted instruction of the skills being assessed. Second, assessment systems should be designed so that relevant results should be available to users at different levels (perhaps quite different information at different levels), from the test-taker to their instructor(s) and then beyond, varying of course in the level of specificity required by each stakeholder. Of course, this would need to be designed within suitable cost and usage limits—the desire for efficiency that leads to the use of matrix item sampling precludes the use of the resulting estimates at finer levels of the system (e.g., in a matrix sample design, individual student results may not be useable due to the small number of items that any one student selects from a particular construct). A reporting system that is transparent at all levels will lead to increased feedback and understanding of student development (Hattie 2009). Finally, end users (e.g., teachers, school administrators, state education leaders, etc.) must be given training in ways of turning results into educational progress. Moreover, test takers benefit if both the assessment and reporting systems can be modified to be “in-synch” with curricula, rather than taking time away from instruction.

At the student level, assessments should dovetail with instruction and provide formative and summative information that allows teachers, instructors, or mentors to better understand the strengths and weaknesses of students’ learning and of teaching practices. The reporting of results from such assessments is crucial to the implementation of remediation or tutoring plans, as the results characterize the extent to which a test taker has demonstrated proficiency in the focal area(s) of the assessment. So the form the results take must guide the decision process as to the best course of action to aid a test taker.

An example of a result report from the DCM example described above is shown in Fig. 3.15.

To maximize the effectiveness of testing, results must be available to users at all levels. We expect the biggest impact to be felt at the most direct levels of users: the students and teachers. At the student level, self-directed students can be helped to understand what skills they are weak in and study accordingly, perhaps with help from their parents or guardians. Similarly, at the teacher level, teachers can examine trends in their class and focus their instruction on areas where students are assessed as deficient. Furthermore, teachers can identify students in need of extra assistance and can, if resources permit, assign tutors to such students.



Diagnostic Scoring Report

Student Name: Daphne

Review Your Answers

Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Your Answer	✓	✓	✓	✓	a	c	✓	c	d	✓	✓	✓	c	✓	d	a	✓	b	a	a	d	c	b	a	c
Correct Answer	d	a	b	d	d	a	b	d	a	c	a	b	d	c	a	d	a	c	b	d	a	a	a	d	b
Difficulty	e	e	m	m	m	m	h	h	h	m	e	e	m	m	m	h	m	m	h	h	h	h	h	h	h

Score

You correctly answered 10 out of 25 questions.

Easy: 4/4; Medium: 5/10; Hard: 1/11

Guide

✓ - Correct answer; o - Omitted answer

e - Easy; m - Medium; h - Hard

Improve Your Skills

Example Questions
3, 14, 2, 17, 19, 23, 9
3, 12, 13, 5, 2, 17, 18, 16, 24, 7
11, 15, 1, 8, 18
22, 20, 10, 11, 5, 6, 18, 25
4, 14, 20, 12, 5, 19, 9

Fig. 3.15 Example of score report from diagnostic classification model analysis (Adapted from Rupp et al. 2010)

Of course, there must also be reporting of results at levels beyond the student and teacher. Schools, districts, reporting regions (states or provinces), and nations all can report such results so that underperforming areas can be identified and addressed accordingly. By understanding the unique dynamics likely to be in place at each level, more efficient systems of remediation can be implemented to help students learn and grow. An example of the use of profile graphs in the reporting of results to different stakeholders in a national evaluation process is presented at the end of this section.

In this example, Chile's National Teacher Evaluation Program reports the results of the evaluation not only to the teacher but also to the corresponding municipality, in order to inform the planning of the professional development courses in the region. This information is disaggregated in several dimensions and compared with the national trend in order to help identify the areas that require more attention.

Designers of tests and assessments need to train end users to be well versed in how to properly harness the information being presented in reports. Although it may seem to need no mentioning, a lack of understanding of test results makes the testing process pointless, as we feel it is a key to understanding the knowledge states of students. Well-informed users, such as teachers and administrators, can turn results into action, assisting students who are in need of remediation or designing challenging exercises for students who are excelling in a content area. Without training, we fear that any benefits of testing may not be realized and that exercise would simply amount to time spent away from instruction, making the learning process much less efficient.

Reporting to Users: An Example from Chile's National Teacher Evaluation

The reports used by the Chilean National Teacher Evaluation (NTE) system (http://www.docentemas.cl/) offers an example of reporting results in terms of profiles in order to provide formative information based on the results of the assessment process.

The NTE is a mandatory process that must be completed by all teachers working in the public school system. The evaluation addresses eight dimensions of teacher performance: Content Organization, Quality of Class Activities, Quality of Assessment Instruments, Use of Assessment Results, Pedagogical Analysis, Class Climate, Class Structure, and Pedagogical Interaction. In each of these eight dimensions, the teachers are evaluated in terms of four different proficiency levels: Unsatisfactory, Basic, Competent, and Outstanding. The evaluation comprises four assessment instruments: a self-evaluation, a peer assessment, a supervisor assessment, and a portfolio assessment completed by the teacher, which includes written products and a recorded lesson (DocenteMas 2009).

Based on their overall results, the teachers are also assigned a general proficiency level. On the one hand, teachers whose overall proficiency level is unsatisfactory or basic are offered professional training programs to improve their performance, while on the other, teachers whose overall proficiency is either competent or outstanding become eligible for an economic incentive.

The NTE provides reports of the results to a variety of users and stakeholders on different roles in the educational system. Among these, arguably the most important reports are the individual reports given to teachers and the summary scores given to every municipality. The relevance of the latter is that the decisions on the contents and structure of professional development that is offered to the teachers are determined at the municipality level.

To provide actionable information to both the teachers and the municipalities, it is critical to go beyond the report of the overall category and present detailed results on each of the eight dimensions that are assessed. In order to do so, the NTE provides profile reports both to municipalities and teachers. Figures 3.16 and 3.17 present examples of the types of graphs used by the NTE in its reports.

Figure 3.16 shows a sample graph used in the reports to municipalities. The graph presents three profiles of the eight dimensions: (1) the average results for the national sample of teachers, (2) the average results for all the teachers in that municipality, and (3) the average results for the teachers whose results locate them in the lower performance categories, namely, basic and unsatisfactory (B + U). This information can then be used by each municipality in the creation of their professional development programs, ideally allowing them to focus on the areas that appear to be most problematic for their teachers.

Additionally, all teachers receive detailed reports about their results of each of the assessment instruments; Fig. 3.17 presents an example of the type of summary graph used in the feedback reports for individual teachers. The profile report for each teacher does not present several profiles, only the one corresponding to the particular teacher; however, in this case, each of the eight assessment dimensions is associated

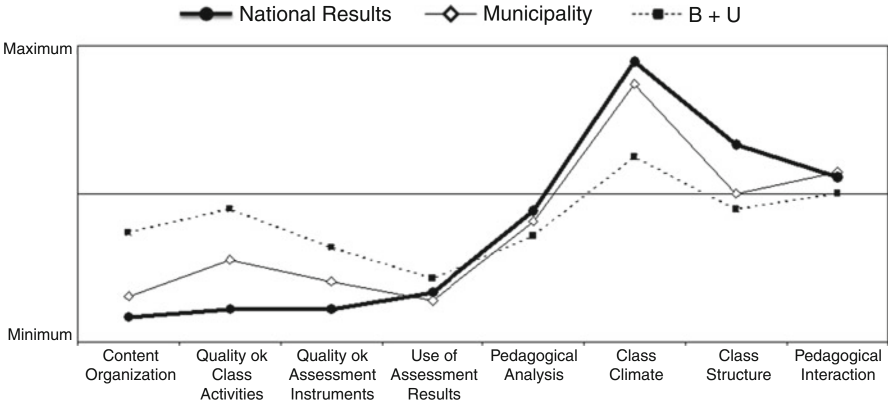


Fig. 3.16 Profile report at the municipality level

PORTFOLIO RESULTS

	Content Organization	Quality of Class Activities	Quality of Assessment Instruments	Use of Assessment Results	Pedagogical Analysis	Class Climate	Class Structure	Pedagogical Interaction
Outstanding		●						
Competent	●					●		●
Basic				●	●		●	
Unsatisfactory			●					

Fig. 3.17 Profile report at the teacher level

with a performance level. This summary profile is complemented by a written report that elaborates the description of their performance level in each dimension.

The use of profile reports is a simple way of conveying information in the context of multidimensional measures, and it allows going beyond the classification of students or teachers in terms of a summary score.

Issues in the Assessment of Twenty-First Century Skills

Generality Versus Context Specificity

When defining constructs for measurement, a key question that can arise is the degree to which any particular context will influence measures of the construct. For instance, in assessments of vocabulary for reading, is the context of a passage selection important? For two respondents who may not fundamentally differ on the overall construct, will different instructional routes have led to different results when the testing is set in different contexts? What aspects of context may imply multidimensionality in the construct?

In traditional measurement, this has been a long-standing concern. Checks for multidimensionality and the examination of evidence for both content validity and internal structure validity are reflections of the issue. They may be addressed from a sampling perspective, by considering sufficiently representative sampling over the potential contexts to provide a reasonable overall measure of the construct.

However, with twenty-first century skills, the context can be quite distal from the construct. For instance, if communication is considered, it can be measured within numerous subject matter areas. Communication skills in mathematics, involving a quantitative symbolic system, representations of data patterns, and so forth, are different from communication skills in, for instance, second-language acquisition, where mediation and meaning-making must occur across languages. On the other hand, some underlying aspects may be the same for communication across these contexts—it may be important to secure the listener's or audience's attention, monitor for understanding, and employ multiple avenues for understanding, across contexts.

At this point in the development of robust measures of twenty-first century skills, there may be more questions than answers to questions of context. Some educators see context as amounting to an insurmountable barrier to measurement. They may claim that the item-specific variability is so high that sound generalizations are not possible. Or they may go further and believe that there is no such thing as a general construct, just specific performances for specific contexts as measured by specific items.

The means of addressing these concerns will necessarily vary from context to context, with some proving more amenable to generalization than others—that is, some contexts may be more specific than others. To some extent, this may have to do with the "grain size" of the generalization and the purpose to which it is put. For instance, a very fine-grained cognitive diagnostic analysis of approaches to problem-solving for a topic such as "two-variable equation systems in beginning Algebra" may or may not generalize across contexts; the question of the stability of a student's approach to "quantitative reasoning across subject matter areas" is likely to be a different question, with perhaps a different answer.

The exploration of context specificity versus context generality is an exciting area of investigation, and it should not be seen as a barrier so much as an opportunity to explore and advance understanding. Some key questions to consider include whether the context may alter proficiency estimates for the construct, such as have been described above. This introduces questions of stability of the measures and also calls for investigations of multidimensionality as these contexts are explored. Opportunities seem ripe for investigating commonalities of constructs across contexts and divergences. Numerous methodological tools are available to consider the stability of constructs across contexts, and now may be an excellent time to do this, considering the nature of twenty-first century needs for skills and knowledge across contexts.

Another important but rather different aspect of the context to consider is the purpose of the assessment. A construct may not be defined in exactly the same fashion in all types of use. For instance, in a formative or classroom-based setting, the context may specifically include local or regional information and issues or topics relevant to a particular community. This tactic may reinforce situated cognition or underscore important learning goals in the local context, and be important to assess for teaching and learning purposes. However, it may not be helpful to include this as a context in a larger-scale summative assessment that is intended to reach beyond the local setting.

Determining what to assess and how to assess it, whether to focus on generalized learning goals or domain-specific knowledge, and the implications of these choices have been a challenge for educators for many years. Taxonomies such as Bloom's Taxonomy of Educational Objectives (Bloom 1956), Haladyna's Cognitive Operations Dimensions (Haladyna 1994), and the Structure of the Observed Learning Outcome (SOLO) Taxonomy (Biggs and Collis 1982) are among many attempts to concretely identify generalizable frameworks. Such frameworks can be very helpful in defining constructs. As Wilson (2009, p. 718) indicates:

> *… as learning situations vary, and their goals and philosophical underpinnings take different forms, a "one-size-fits-all" development assessment approach rarely satisfies educational needs.*

A third perspective in the issue of context specificity can be found in the research on expert–novice differences. In its review of the research, the *How People Learn* report emphasizes that expertise is not domain-general, but on the contrary, is related to *contexts of applicability,* indicating that knowledge is conditional on a set of circumstances (NRC 2000) that can be so cryptic that it hinders the transfer of learning, amounting to subject chauvinism. This is surely not a separate issue but the dimensionality issue that was raised earlier.

The notion of expertise can be used to shed some light about the role of context in the evolution of a skill. As mentioned above, the "expert" in a competency offers a natural upper level within which we can focus, but in addition to informing us about this upper stage, research in the field of expertise can provide insight about the evolution of these skills.

From the perspective of this research tradition, the contextual nature of knowledge goes beyond the issue of selection of relevant domains associated with a particular construct, calling attention to the question of *when* certain knowledge becomes relevant. The NRC report had this to say on this issue:

> The concept of conditionalized knowledge has implications for the design of curriculum, instruction, and assessment practices that promote effective learning. Many forms of curricula and instruction do not help students conditionalize their knowledge: "Textbooks are much more explicit in enunciating the laws of mathematics or of nature than in saying anything about when these laws may be useful in solving problems" (Simon 1980:92). It is left largely to students to generate the condition-action pairs required for solving novel problems. (NRC 2000, p 43)

The challenges associated with the effective application of knowledge have also been recognized by advocates of a domain-general approach. In his discussion of the issues that must be faced if one is to attempt to teach general skills, Hayes (1985) indicates (a) that general skills require immense knowledge covering all the potential contexts in which the skill will be applied and (b) that general strategies need to deal with the problems of being appropriately identified and transferred to the context of particular problems.

It is important to note that the relevance of these debates goes beyond the purely theoretical. The adoption of a domain general versus a context specific approach will have practical implications in the description of learning targets, progress variables, levels of achievement, and learning performances. Different options will affect the grain size of the operational definitions and will determine the specificity of the characterization of the products and actions that are to be considered as evidence of performance, therefore circumscribing the domains in which is possible to make inferences.

Large-Scale and Classroom Assessments

The kinds of inferences that we intend to make will influence the evidence that will be collected. In this sense, a crucial area that requires definition is clarifying who are the intended users and, tied to that, the levels of analysis and reporting that need to be addressed. The range of intended users and stakeholders will delineate the scope of the project, the characteristics of the data that needs be collected, and consequently the methodological challenges that must be met.

A direct consequence of determining the intended user is the realization that what constitutes useful information for decision making will vary widely between its recipients, for example, teachers and government officials. In other words, the kind of and level of detail required to support student learning differ in the classroom from that at the policy level. At the same time, it is necessary to keep in mind that making reliable inferences about each student in a classroom requires considerably more information than making inferences about the class as a whole; in other words, reaching an adequate level of precision to support inferences demands considerably more data when individuals are being discussed.

If the assessment of twenty-first century skills is to address the concerns of both teachers and governments, it is necessary to (a) determine what is needed in the classroom, what is good for formative assessment and what is helpful for teachers, (b) determine what is good in terms of large-scale assessment, and (c) achieve consistency between those levels without imposing the restrictions of one upon the other. We should not pursue large-scale assessment without establishing classroom perspectives and assessments for the same variables.

The imposition of the requirements of large-scale assessments upon the classroom can have negative consequences, such as the creation of de facto curricula focused only on the elements present in standardized instruments. In the case of the assessment of twenty-first century skills, this particular problem raises two potential risks. The first of these is related to the practical consequences of the inclusion of new sets of competencies in classrooms, which could overwhelm teachers and students with additional testing demands and/or modify their practices in counterproductive ways. The second one is related to the potential restriction of the twenty-first century skills by the instruments used to assess them; reliance on large-scale assessments may distort the enactment of curriculum related to the twenty-first century skills, with negative impacts on both the development of the skills and the validity of the assessments.

A potential solution for the articulation of these different levels, ranging on a continuum from classroom tasks to large-scale assessments, will probably include the use of unobtrusive artifacts and proxies that allow the collection of information from the classrooms in order to inform large-scale assessments. The exploration of alternative and novel information sources that could provide valid data without the need to use additional measurement instruments could be very convenient (e.g., nonintrusive), due to their indirect nature. However, two issues that need to be considered in the use of proxies are (a) the trade-off between the specificity of certain tasks and the ability to make inferences to different contexts associated with the constructs of interest, and (b) their credibility for the kind of inference that will be drawn from them. For example, they might be useful and widely accepted at the classroom level, but administrators and policy makers could resist their interpretation as large-scale indicators.

At the same time, users of these kinds of indicators must confront the fact that the definitions of daily activities as forms of assessment can change the nature of the practices being measured. For example, if the number of exchanged emails is used now as an indicator of engagement in a community, this very definition of number of emails as an input to evaluation could alter the underlying dynamic, generating issues such as playing the system, prompting students to deliberately send more emails in order to boost the engagement indicator.

To summarize, a clear methodological challenge arising from defining the range of intended users is the articulation between the different levels that they represent. This challenge takes many forms. On the one hand, it raises questions about our capacity to generate new forms of nonintrusive assessment that can be used to inform large-scale assessments without perturbing classroom dynamics. On the other hand, it raises questions about how to provide pertinent information flow in the opposite direction: how to use standardized test data to inform classroom practices?

What Can New Advances in Technology Bring to Assessments?

New advancements in technology can bring a myriad of new functions to assessments. Much has been described in the research literature about dynamic visuals, sound, and user interactivity, as well as adaptivity to individual test takers and near real-time score reporting and feedback in online settings (Bennett et al. 1999; Parshall et al. 2000, 2002, 1996; Scalise and Gifford 2006). These take advantage for assessments of the availability of new types of computer media and innovative approaches to its delivery.

However, a new direction of innovation through technology may be even more fundamentally transforming what is possible through assessment. It has to do with social networking and online collaboration, or so-called "Web 2.0." The evolving paradigm involves several influences. These might be summed up by such concepts as "wisdom of the crowd," personalization, adaptive recommender systems, and "stealth" assessment.

Wisdom of the Crowd

Crowd wisdom is built on "prediction markets," harvesting the collective wisdom of groups of ordinary people in order to make decisions (Giles 2005; Howe 2008). Advocates believe these predictions can sometimes be better than those made by specialists. It is viewed as a way to integrate information more empirically, and forms of this can be seen in the use of historic data in data mining, as well as in forecasting, such as through the use of user ratings and preference votes of various kinds. Wisdom of the crowd as a concept is still quite novel in educational assessment. For instance ERIC, a large repository of educational research, when searched in August 2009, had only two citations that even mentioned wisdom of the crowd, and neither of these concerned ICT. So how can this be useful in education, when the goal is not predicting a vote or making a direct forecast? Educationally, the thought is that, in similar ways to evaluations and user response processes, information can be harvested from "the crowd" to improve offerings and make ratings on individuals, groups, educational approaches, and so forth. Of course, it is easy to do this if it is not clear how the evidence is being interpreted and if decisions are low-stakes for an undefined group of stakeholders. It is harder to obtain "wisdom" that is defensible for high-stake decisions, so we would want to investigate the interaction between different aspects here.

With the rise of social networking, audiences interested and invested in wisdom of the crowd decisions are growing quickly. In the age group 14–40, individuals now spend more time in social networking online than in surfing the Web (Hawkins 2007). Choice, collaboration, and feedback are expected within this community. Also, it is important to realize that social networking emphasizes "friends," or affiliates, rather than hierarchical structures. Loyalty to networks can be high, and engagement is built on the passion of participants, who have actively chosen their networks (Hawkins 2007).

To consider what crowd wisdom is in more formal assessment terms, it may be useful to describe it in terms of the four building blocks of measurement mentioned above: the construct—or the goals and objectives of measurement, the observations themselves that provide assessment evidence, the scoring or outcome space, and the measurement models applied (Wilson 2005). In wisdom of the crowd, it is not so much the observations themselves that change—ultimately, they will often remain individual or group indicators of some measured trait or attribute. New media or interactions may or may not be used, but fundamentally, from the social networking standpoint, what is changing is the comparison of these attributes to what might be considered different types of group norming. This can be done through the ways the attributes are ultimately "scored" or interpreted relative to profiles and group wisdom considerations. It may involve numerous individuals or groups rating one another, and it may even fundamentally redefine the construct in new directions associated with group thinking.

An example to consider in education has been suggested by Lesgold (2009). He describes how one could "imagine asking teachers to test themselves and to reflect on whether what they are teaching could prepare students to do the tasks that companies say validly mirror current and future work." If broadly collected across many teachers, this could be an example of crowd wisdom of teachers, helping to define interpretation of scores and benchmarking, or even the development of new constructs. In many ways, it is not entirely different from such activities as involving teachers in setting bookmark standards, but extends the idea greatly and moves it from a controlled context of preselected respondents to a potentially much broader response base, perhaps built around non-hierarchically organized groups or social networks.

Lesgold suggests another approach that might be seen as tapping crowd wisdom, this time from the business community.

> Would it be worthwhile [he asks,] to develop a survey that apprenticeship program recruiters could fill out for each applicant that provided a quick evaluation of their capabilities, perhaps using the Applied Learning standards as a starting point for developing survey items?…. One can compare current public discussion of No Child Left Behind test results to ads from American car manufacturers that give traditional measures such as time to accelerate to 60 mi/hr, even though the public has figured out that safety indicators and economy indicators may be more important. We need to reframe the public discussion, or schools may get better at doing what they were supposed to do half a century ago but still not serve their students well. (Lesgold 2009, pp. 17–18)

Surveys such as Lesgold describes are currently being used in business contexts to "rate" an individual employee's performance across numerous groups and instances, based on what respondents in the networks say about working with the individual on various constructs. An aggregate rating of crowd wisdom is then compiled, from a variety of what could be counted as twenty-first century skills, such as teamwork, creativity, collaboration, and communication.

Adaptive Recommender Systems

Adaptive recommender systems take the wisdom of the crowds one step farther, using assessment profiles to mediate between information sources and information seekers (Chedrawy and Abidi 2006), employing a variety of methods. The goal is to determine the relevance and utility of any given information in comparison to a user's profile. The profile can include attributes such as needs, interests, attitudes, preferences, demographics, prior user trends, and consumption capacity. The information can take many forms, but is often based on a much broader set of information and the use of various forms of data mining rather than on authoritative, expert, or connoisseur analysis as in some of the semipersonalized ICT products described above.

Stealth Assessment

The term “stealth assessment” simply implies that this type of diagnosis can occur during a learning—or social networking—experience and may not necessarily be specifically identified by the respondent as assessment (so can be thought of as a type of unobtrusive measure). Then instructional decisions can be based on inferences of learners’ current and projected competency states (Shute et al. 2009, 2010). Shute describes inferences—both diagnostic and predictive—handled by Bayesian networks as the measurement model in educational gaming settings. Numerous intelligent tutoring systems also exist that rely on more or less overt educational assessments and employ a variety of measurement models to accumulate evidence for making inferences.

Personalization

Personalization in information and communication technology (ICT) adjusts this “stealth” focus somewhat by specifically including both choice and assessment in the decision-making process for learners. This means that so-called “stealth” decisions and crowd wisdom are unpacked for the user and presented in a way such that choice, or some degree of self-direction, can be introduced.

Personalization was cited by *Wired* magazine as one of the six major trends expected to drive the economy in upcoming years (Kelleher 2006). Data-driven assessments along with self-directed choice or control are becoming common dimensions for personalization in such fields as journalism (Conlan et al. 2006), healthcare (Abidi et al. 2001), and business and entertainment (Chen and Raghavan 2008). Personalized learning has been described as an emerging trend in education as well (Crick 2005; Hartley 2009; Hopkins 2004), for which ICT is often considered one of the promising avenues (Brusilovsky et al. 2006; Miliband 2003). With regard to ICT, the goal of personalized learning has been described as supporting “e-learning content, activities and collaboration, adapted to the specific needs and influenced by specific preferences of the learner and built on sound pedagogic strategies” (Dagger et al. 2005, p. 9). Tools and frameworks are beginning to become available to teachers and instructors for personalizing content for their students in these ways (Conlan et al. 2006; Martinez 2002).

Examples of Types of Measures

Assessment of New Skills

Assessment of the routine skills of reading and math are currently reasonably well developed. However, in the workplace, as Autor et al. (2003) point out, demand for some of the routine cognitive skills that are well covered by existing standardized tests is declining even faster than the demand for routine and nonroutine manual skills. According to Levy and Murnane (2006), the skills for which demand has grown most over the last 30 years are complex communication, and expert thinking and problem-solving, which they estimate to have increased by at least 14% and 8%, respectively.

Assessment of these new skills presents many challenges that have either been ignored completely, or substantially underplayed, in the development of current standardized assessments of the outcomes of K-12 schooling. Perhaps the most significant of these is, in fact, inherent in the definition of the skills.

In all developed countries, the school curriculum has been based on a model of "distillation" from culture, in which valued aspects of culture are identified (Lawton 1970) and collated, and common features are "distilled." In some subjects, particular specifics are retained, for example in history, it is required that students learn about particular episodes in a nation's past (e.g., the civil war in the USA, the second world war in the UK, etc.), and in English language arts, certain canonical texts are prescribed (e.g., Whitman, Shakespeare). However, at a general level, the process of distillation results in a curriculum more marked for generality than for specificities. This is perhaps most prominent in mathematics, where in many countries, students are still expected to calculate the sum of mixed fractions even though real-life contexts in which they might be required to undertake such an activity are pretty rare (except maybe when they have to distribute pizza slices across groups of eaters!). Such a concern with generality also underlies the assessment of reading since it is routinely assumed that a student's ability to correctly infer meaning from a particular passage of grade-appropriate reading material is evidence of their ability to read, despite the fact that there is increasing evidence that effective reading, at least at the higher grades, is as much about understanding the background assumptions of the author as it is about decoding text (Hirsch 2006).

Such an approach to curriculum poses a significant problem for the assessment of twenty-first century skills because of the assumption that these skills will generalize to "real" contexts even though the evidence about the generalizability of the skills in the traditional curriculum is extremely limited.

Typical sets of skills that have been proposed for the label "twenty-first century skills" (e.g., Carnevale et al. 1990) are much less well defined than the skills currently emphasized in school curricula worldwide. Even if the challenge of construct definition is effectively addressed, then because of the nature of the constructs involved, they are likely to require extended periods of assessment. Even in a relatively well-defined and circumscribed domain, such as middle and high school science, it has been found that six tasks are required to reduce the construct-irrelevant variance associated with person by task interactions to an acceptable level (Gao et al. 1994). Given their much more variable nature and the greater variety of inferences that will be made on the basis of the assessment outcomes, the assessment of twenty-first century skills may well require a very large number of tasks—and almost certainly a larger number than is imagined by those advocating their adoption.

Self-Assessment and Peer Assessment

There is evidence that, although peer and self-assessments are usually thought most suitable for formative assessments, they can also effectively support summative inferences, but where high stakes attach to the outcomes, it seems unlikely that this will be the case. Having said this, a large corpus of literature that attests that groups of individuals can show a high degree of consensus about the extent to which particular skills, such as creativity, have been demonstrated in group activities.

Wiliam and Thompson (2007) point out that self and peer assessment are rather narrow notions and are more productively subsumed within the broader ideas of "activating students as owners of their own learning" and "activating students as learning resources for one another," at least where the formative function of assessment is paramount. In a sense, accurate peer and self-assessment can then become a measure of certain types of metacognition.

Sadler (1989) says:

> The indispensable conditions for improvement are that the student comes to hold a concept of quality roughly similar to that held by the teacher, is continuously able to monitor the quality of what is being produced during the act of production itself, and has a repertoire of alternative moves or strategies from which to draw at any given point. (p. 121)

This indicates again that adequate construct definition will be essential in the operationalization and promulgation of twenty-first century skills.

Creativity/Problem-Solving

Definitions of creativity and problem-solving also have some embedded dilemmas for measurement. Mayer (1983) says:

> Although they express the terms differently, most psychologists agree that a problem has certain characteristics:
>
> *Givens*—The problem begins in a certain state with certain conditions, objects, pieces of information, and so forth being present at the onset of the work on the problem.
>
> *Goals*—The desired or terminal state of the problem is the goal state, and thinking is required to transform the problem from the given state to the goal state.
>
> *Obstacles*—The thinker has at his or her disposal certain ways to change the given state or the goal state of the problem. The thinker, however, does not already know the correct answer; that is, the correct sequence of behaviours that will solve the problem is not immediately obvious. (p. 4)

The difficulty with this definition is that what may be a problem for one student is simply an exercise for another because of the availability of a standard algorithm. For example, finding two numbers that have a sum of 10 and a product of 20 can result in worthwhile "trial and improvement" strategies, but for a student who knows how to resolve the two equations into a single quadratic equation and also knows the formula for finding the roots of a quadratic equation, it is merely an exercise. Whether something is a problem therefore depends on the knowledge state of the individual.

For some authors, creativity is just a special kind of problem-solving. Newell et al. (1958) defined creativity as a special class of problem-solving characterized by novelty. Carnevale et al. (1990) define creativity as "the ability to use different modes of thought to generate new and dynamic ideas and solutions," while Robinson defines creativity as "the process of having original ideas that have value" (Robinson 2009). Treffinger (1996) and Aleinikov et al. (2000) each offer over 100 different definitions of creativity from the literature. Few, if any, of these definitions are sufficiently precise to support the precise definition of constructs required for the design of assessments.

Whether creativity can be assessed is a matter of much debate, compounded, as mentioned above, by the lack of a clear definition of what, exactly, it is. The Center for Creative Learning (2007) provides an index of 72 tests of creativity, but few validity studies exist, and even fewer that would support the use of the principles of evidence-centered design.

Group Measures

Threaded throughout this chapter have been examples of twenty-first century skills playing out in group contexts. Even in approaches such as personalized learning, group interaction and teamwork are fundamental; personalized learning does not mean strictly individualized instruction, with each student learning on their own (Miliband 2003). On the contrary, the twenty-first century view of such learning opportunities promotes teamwork and collaboration and supports learning and student work in classes and groups. The call for personalization (see above) includes options rich with the possibility of human interaction, with some commentators suggesting that proficiency at the higher levels of cognitive functioning on Bloom's taxonomy encourages Web 2.0 social activities, such as information sharing and interaction (Wiley 2008). Within the context of interaction, personalization refers to "rigorous determination to ensure that each student's needs are assessed, talents developed, interests spurred and their potential fulfilled" (Miliband 2003, p. 228). The idea of a zone of proximal development (ZPD) in personalization is that the learning environment presents opportunities for each student to build their own understanding within a context that affords both group interaction and individual challenge.

Thus, we can ask: What methodological approaches can be used for assessment within group settings? Much regarding this remains to be explored in depth as an upcoming agenda for researchers in the measurement field, but a small set of eight possibilities is shown below. Other additional approaches might be postulated, and those listed here are given as some examples only and are not intended to be exhaustive but only suggestive. They are sorted as innovations in terms of which of the four fundamental building blocks of measurement they address: the construct—or the goals and objectives of measurement, the observations themselves that provide assessment evidence, the scoring or outcome space, and any measurement models applied (Wilson 2005).

Construct:

1. Changing views of knowledge suggest that a reinterpretation of at least some of what twenty-first century skills mean might be helpful at the construct level, for instance, defining the construct functionally as *only* those aspects of it that operate within a group. This could be the case for group leadership, group facilitation, and so forth. The group member is then scored on outcomes of her or her role on this construct within the group. Sampling over several groups would probably give a more representational score.
2. Use wisdom of the crowd and feedback from all group members within the group to provide information on each individual's contribution on the various indicators of interest, within a series of groups (an example of this is the business environment use of employee success within group settings).

Observation:

3. Use the much improved possibilities of data density (discussed above), along with representational sampling techniques, to aggregate group performance for the individual across many groups and over multiple contexts and conditions.
4. Collect individual indicators while participating in a group, and in advance of group work on each particular indicator. An example of this is "prediction" indices, where each member of a group "predicts" their expected outcome prior to reflection and work by the group (Gifford 2001).

Outcome space:

5. Work products are scored on two scales, one for individual performance and one for group performance—individual scores can be collected by preestablishing a "role" within the task for each individual, by the submission of a separate portion of the work product, or by submission of duplicate work products (for instance, lab groups with same results, different laboratory report write-ups for each individual).
6. Groups are strategically constructed to consist of peers (with known different abilities) working together on separate constructs, and the more able peer does scoring on each construct. For example, team a German Language Learner who is an English native language speaker with an English Language Learner who is a German native language speaker. They work together synchronously online for written and oral language, each communicating only in the language they are learning. They are then either scored by the more able peer in each language, or the more able peer is required to answer questions showing they understood what was being communicated in their native language, indicating the success of the learning communicator.

Measurement model:

7. Individual performances are collected across numerous instances within more stable groups. A measurement model with a group facet parameter adjusts indicators recorded in several group experiences, similar to the operation of Testlet Response Theory which uses an item response model with a testlet "facet" (Wainer et al. 2006; Wang and Wilson 2005).
8. Both individual and group indicators are collected and used to score a student on the same construct using item response models and construct mapping (Wilson 2005). Fit statistics are used to indicate when an individual has an erratic performance between the two conditions, and one then applies closer examination for the less fitting students.

Biometrics

Much of the discussion on assessments in this chapter has been in terms of answers to questions, tasks, or larger activities that generate responses to different kinds of indicators. Another approach, biometrics, is more implicit and involves tracking actual physical actions.

Biometrics is the science and technology of measuring and statistically analyzing biological data. The derivation, expression, and interpretation of biometric sample quality scores and data are summarized in standards of the International Organization for Standardization (2009) and refers primarily to biometrics for establishing identity, such as fingerprints, voice data, DNA data, Webcam monitoring, and retinal or iris scans. Such physiological characteristics that do not change often can be used for identification and authentication. They have been tried out for use in some settings of high-stakes assessment to authenticate and then to monitor the identity of a respondent during an examination, for instance, in the absence of proctoring (Hernández et al. 2008).

Here we are perhaps more concerned with biometrics as behavioral characteristics that may be a reflection of a user response pattern related to a construct of interest to measure. These include keystroke analysis, timed response rates, speech patterns, haptics (or kinesthetic movements), eye tracking, and other approaches to understanding a respondent's behaviors (Frazier et al. 2004).

To take one example, in keystroke dynamics, the duration of each stroke, lag time between strokes, error rate, and force are all aspects of biometrics that can be measured. These might be useful to understand student proficiency on technology standards that involve keyed interfaces, or if a construct assumed something about these factors. Response rates that are either too fast or too slow may indicate interesting and relevant user characteristic. For instance, in test effort, keystroke response rates that are too fast—faster than even a proficient user could respond on the item—have been used to detect instances of underperformance on test effort in computer adaptive testing (Wise and DeMars 2006; Wise and Kong 2005).

Another area of biometrics, this time involving haptics, or movement, is gait technology, which describes a person's walk, run, or other types of motion of the leg. Similar technologies consider other bodily motions. These might be used in physical education, analysis for repeated motion injury for student athletes, or to optimize a physical performance skill. For instance, posture assessments in a progress monitoring system of functional movement assessment for injury prevention used for student athletes at one US university cut injury outage rates for the men's intercollegiate basketball team from 23% during the basketball season to less than 1% in a single year.

Sensors and performance assessments such as devices with embedded sensors that allow the computer to see, hear, and interpret users' actions are also being tried in areas such as second language acquisition, through approaches known as ubiquitous (ever-present) computing (Gellersen 1999).

Eye tracking is another area beginning to receive some attention. Here, assessments of what students are focusing upon in the computer screen interface may yield information about their problem-solving approaches and proficiency. If eye tracking shows focus on superficial elements of a representation or data presentation, this might show a less efficient or productive problem-solving process, compared with earlier and more prolonged focus on the more important elements. Such assessments might be used in cognitive diagnosers to suggest possible hints or interventions for learners (Pirolli 2007).

In simulations, for example, it has been shown that anything in motion draws the student's attention first; but, if the simulation simply demonstrates the motion of an object, students rarely develop new ideas or insights (Adams et al. 2008). In these cases, many students accept what they are seeing as a transmitted fact, but are not often seen attempting to understand the meaning of the animation. However, by combining eye tracking with personalization processes that allow user control over the simulations:

> when students see an animated motion instantly change in response to their self-directed interaction with the simulation, new ideas form and they begin to make connections. Students create their own questions based on what they see the simulation do. With these questions in mind, they begin to investigate the simulation in an attempt to make sense of the information it provides. In this way, students answer their own questions and create connections between the information provided by the simulation and their previous knowledge. (Adams et al. 2008, p. 405)

Conclusion

By now it should be clear to the reader that we are not presuming to offer answers to all of the methodological issues regarding twenty-first century skills that we discuss in this chapter. Instead, we have taken the opportunity to raise questions and seek new perspectives on these issues. In concert with this approach, we end the chapter with a set of challenges that we see as being important for the future of both research and development in this area. We do not claim that these are the only ones that are worth investigating (indeed, we have mentioned many more in the pages above). Nor do we claim these are the most important under all circumstances. Effectively, what we have done is looked back over the six major sections of the chapter and selected one challenge from each section—we see this as a useful way to sample across the potential range of issues and help those who are attempting to work in this area to be prepared for some of the important questions they will face. The challenges are as follows:

How can one distinguish the role of context from that of the underlying cognitive construct in assessing twenty-first century skills? Or, should one?

Will the creation of new types of items that are enabled by computers and networks change the constructs that are being measured? Is it a problem if they do?

What is the balance of advantages and disadvantages of computerized scoring for helping teachers to improve their instruction? Will there be times when it is better not to offer this service, even if it is available?

With the increased availability of data streams from new assessment modes, will there be the same need for well-constructed outcome spaces as in prior eras?

How will we know how to choose between treating the assessment as a competitive situation (requiring us to ignore information about the respondents beyond their performances on the assessment), as opposed to a "measurement" situation, where we would want to use all relevant ancillary information? Or should both be reported?

Can we use new technologies and new ways of thinking of assessments to gain more information from the classroom without overwhelming the classroom with more assessments?

What is the right mix of crowd wisdom and traditional validity information for twenty-first century skills?

How can we make the data available in State-mandated tests actionable in the classroom, and how can we make data that originates in the classroom environment useful to state accountability systems?

How can we create assessments for twenty-first century skills that are activators of students' own learning?

The list above is one that we hope will be helpful to people who are developing assessments for twenty-first century skills. The questions are intended to provoke the sorts of debates that should be had about any new types of assessments (and, of course, there should be similar debates about the traditional sorts of tests also).

We will be considering these questions, as well as the others we have mentioned in the pages above, as we proceed to the next phase of our project's agenda—the construction of assessments of some exemplary twenty-first century skills. No doubt, we will have the chance then to report back on some of the answers that we come up with as we carry out this development task, and we will also have the opportunity to take (or ignore) our own advice above.

Acknowledgement We thank the members of the Working Group who have contributed ideas and made suggestions in support of the writing of this paper, in particular, Chris Dede and his group at Harvard, John Hattie, Detlev Leutner, André Rupp, and Hans Wagemaker.

Annex: Assessment Design Approaches

Evidence-Centered Design

Design, in general, is a prospective activity; it is an evolving plan for creating an object with desired functionality or esthetic value. It is prospective because it takes place prior to the creation of the object. That is, a design and the resulting object are two different things (Mitchell 1990, pp. 37–38):

> when we describe the forms of buildings we refer to extant constructions of physical materials in physical space, but when we describe designs we make claims about something else—constructions of imagination. More precisely, we refer to some sort of model—a drawing, physical scale model, structure of information in computer memory, or even a mental model—rather than to a real building.

The idea of design is, of course, equally applicable to assessments, and Mitchell's distinction just noted is equally applicable. The *design of an assessment* and the resulting *assessment-as-implemented* are different entities. Under the best of circumstance, the design is sound and the resulting assessment satisfies the design, as evidenced empirically through the administration of the assessment. Under less ideal circumstances, the design may not be sound—in which case only by a miracle will the resulting assessment be sound or useful—or the implementation of the assessment is less than ideal. In short, merely using a design process in no way guarantees that the resulting assessment will be satisfactory, but it would be foolish to implement an assessment without a thorough design effort as a preamble.

An approach to assessment design that is gaining momentum is ECD, evidence-centered design (Mislevy et al. 2003b). The approach is based on the idea that the design of an assessment can be facilitated or optimized by taking into consideration the *argument* we wish to make in support of the *proposed* score interpretation or inference from the assessment. In its barest form a proposed score interpretation takes the following form: *Given that the students has obtained score X, it follows that the student knows and can do Y.*

There is no reason for anyone to accept such an assertion at face value. It would be sensible to expect an elaboration of the reasons, an argument, before we accept the conclusion or, if necessary, challenge it. A Toulmian argument, whereby the reasons for the above interpretation are explicated and potential counterarguments are addressed, is at the heart of ECD. ECD focuses on that argument primarily, up to the test score level, by explicating what the intended conclusions or inferences based on scores will be, and, given those inferences as the goal of the assessment, determine the observation of student performance that would leads us to those conclusions. Such an approach is in line with current thinking about validation, where a distinction is made between (1) a validity argument, the supporting reasoning for a particular score interpretation, and (2) the appraisal of that argument. ECD turns the validity argument on its head to find out what needs to be the case, what must be true of the assessment—what should the design of the assessment be—so that the score interpretations that we would like to reach in the end will have a better chance of being supported.

For example, suppose we are developing an assessment to characterize student's mastery of information technology. If we wish to reach conclusions about this, we need to carefully define what we mean by "command of information technology," including what behavior on the part of students would convince us that they have acquired mastery. With that definition in hand, we can then proceed to devise a series of tasks that will elicit student behavior or performance indicative of different levels of command of information technology, as we have defined it. Then, as the assessment is implemented, trials need to be conducted to verify that, indeed, the items produced according to the design elicit the evidence that will be needed to support that interpretation.

Approaching assessment development this way means that we have well-defined expectations of what the data from the assessment will look like. For example, what the difficulty of the items will be, how strongly they will intercorrelate, and how the scores will relate to other test scores and background variables. Those expectations are informed by the knowledge about student learning and developmental considerations that were the basis of the design of the assessment; if they are not met, there will be work to be done to find out where the design is lacking or whether the theoretical information used in the design was inadequate.

The process of reconciling design expectations with empirical reality parallels the scientific method's emphasis on hypothesis testing aided by suitable experimental designs. It should be pointed out, however, that an argument based solely on positive confirmatory evidence is not sufficiently compelling. Ruling out alternative interpretations of positive confirmatory evidence would add considerable weight to an argument, as would a failed attempt to challenge the argument. Such challenges can take a variety of forms in an assessment context. For example, Loevinger (1957) argued that items that explicitly aim to measure a different construct should be included, at least experimentally, to ensure that performance in those items is not explained equally well by the postulated construct.

ECD is highly *prospective* about the process for implementing the assessment so that the desired score interpretations can be supported in the end. Essentially, ECD prescribes an order of design events. *First*, the purpose of the assessment needs to be explicated to make it clear what sort of inferences need to be drawn from performance on the test. Once those target inferences are enumerated, the *second* step is to identify the types of evidence needed to support them. Finally, the *third* step is to conceive of means of eliciting the evidence needed to support the target inferences. These three steps are associated with corresponding models: a student model, an evidence model, and a series of task models. Note that, according to ECD, task models, from which items would be produced, are the last to be formulated. This is an important design principle, especially since when undertaking the development of an assessment, there is a strong temptation to "start writing items" before we have a good grasp of what the goals of the assessment are. Writing items without first having identified the target inferences, and the evidence required to support them, risks producing many items that are not optimal or even failing to produce the items that are needed to support score interpretation (see e.g., Pellegrino et al. (1999), Chap. 5). For example, producing overly hard or easy items may be suboptimal if decisions or inferences are desired for students having a broad range of proficiency. Under the best of circumstances, starting to write items before we have a firm conception of the goals of the assessment leads to many wasted items that, in the end, do not fit well into the assessment. Under the worst of circumstances, producing items in this manner can permanently hobble the effectiveness of an assessment because we have to make do with the items that are available.

The importance of a design perspective has grown as a result of the shift to so-called standards-based reporting. Standards-based reporting evolved from earlier efforts at criterion-referenced testing (Glaser 1963) intended to attach a specific interpretations to test scores, especially scores that would define different levels of achievement. Since the early 1990s, the National Assessment of Educational Progress (NAEP) in the USA has relied on achievement levels (Bourque 2009). In the USA, tests oriented to inform accountability decisions have followed in NAEP's footsteps in reporting scores in terms of achievement or performance levels. This, however, does not imply that achievement levels are defined equivalently (Braun and Qian 2007) in different jurisdictions. While it is true that the definition of achievement levels need not, for legitimate policy reasons, be equivalent across jurisdictions in practice in the USA, there has not been a good accounting of the variability across states. A likely reason is that the achievement levels are defined by cutscores that are typically arrived at by an expert panel *after* the assessment has been implemented (Bejar et al. 2007). However, unless the achievement levels have been defined as part of the design effort, rather than leaving them to be based on the assessment as implemented, there is a good chance that there will be a lack of alignment between the intended achievement levels and the levels that emerge from the cutscore setting process. The cutscore setting panel has the duty to produce the most sensible cutscores it can. However, if the assessment was developed without these cutscores in mind, the panel will still need to produce a set of cutscores to fit the assessment *as it exists*. The fact that the panel is comprised of subject matter experts cannot possibly compensate for an assessment that was not designed to specifically support the desired inferences.

Whether the assessment outcomes are achievement level or scores, an important further consideration is the temporal span assumed by the assessment. In a K-12 context, the assessment is focused on a single grade, and, typically, the assessment is administered toward the end of the year. A drawback of a single end-of-year assessment is that there is not an opportunity to utilize the assessment information to improve student achievement, at least not directly (Stiggins 2002). An alternative is to distribute assessments during the year (Bennett and Gitomer 2009); a major advantage of this is the opportunity it gives to act upon the assessment results that occur earlier in the year. Some subjects, notably mathematics and the language arts, can extend over several years, and the yearly end-of-year assessments could be viewed interim assessments. Consider first the simpler case where instruction is completed within a year and there is an end-of-year assessment. In this case, achievement levels can be unambiguously defined as the levels of knowledge expected after 1 year of instruction. For subjects that require a multiyear sequence or for subjects that distribute the assessment across several measurement occasions within a year,

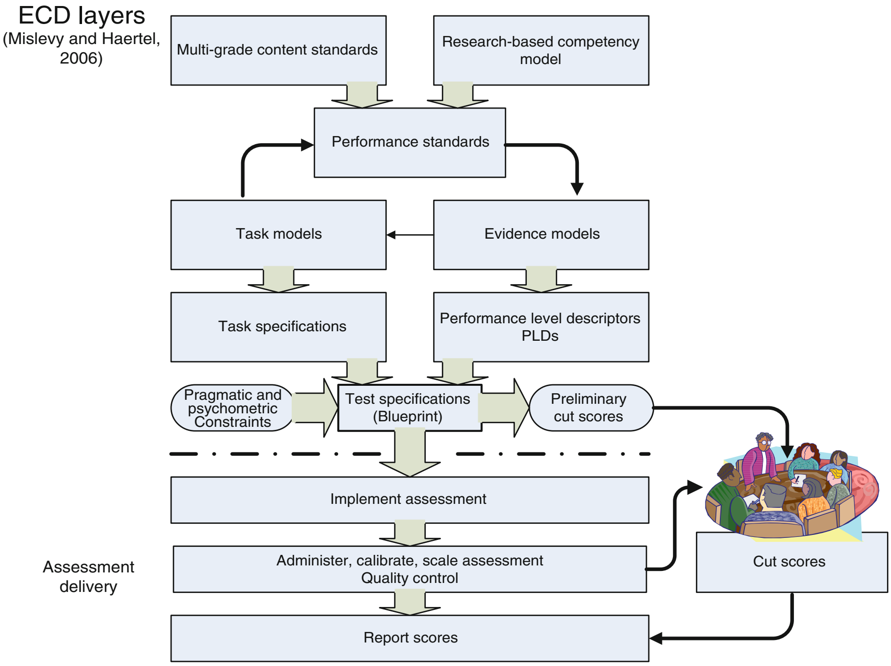

ECD layers
(Mislevy and Haertel, 2006)
Multi-grade content standards
Research-based competency model
Performance standards
Task models
Evidence models
Task specifications
Performance level descriptors PLDs
Pragmatic and psychometric Constraints
Test specifications (Blueprint)
Preliminary cut scores
Implement assessment
Assessment delivery
Administer, calibrate, scale assessment Quality control
Cut scores
Report scores

Fig. 3.18 The ECD framework

at least two approaches are available. One of these defines the achievement levels in a bottom-up fashion. The achievement levels for the first measurement occasion are defined first, followed by the definitions for subsequent measurement occasions. So long as the process is carried out in a coordinated fashion, the resulting sets of achievement levels should exhibit what has been called coherence (Wilson 2004). The alternative approach is top-down; in this case, the achievement levels at the terminal point of instruction are defined first. For example, in the USA, it is common to define so-called “exit criteria” for mathematics and language art subjects that, in principle, define what students should have learned by, say, Grade 10. With those exit definitions at hand, it is possible to work backwards and define achievement levels for earlier measurement occasions in a coherent manner.

Operationalization Issues

The foregoing considerations are some of the critical information in determining achievement levels, which, according to Fig. 3.18, are the foundation on which the assessment rests, along with background knowledge about student learning and developmental considerations. For clarity, Fig. 3.18 outlines the “work flow” for assessment at one point in time, but in reality, at least for some subject matters, the design of "an" assessment really entails the simultaneous design of several. That complexity is captured in Fig. 3.18 under *Developmental considerations*; as the figure shows, *achievement levels* are set by those developmental considerations and a *competency model,* which summarizes what we know about how students learn in the domain to be assessed (NRC 2001).

The achievement levels are fairly abstract characterizations of what students are *expected* to achieve. Those expectations need to be recast to make them more concrete, by means of *evidence models* and *task models.* Evidence models spell out the student behavior that would be evidence of having acquired the skills and knowledge called for by the achievement levels. Evidence models are, in turn, specifications for the tasks or items that will actually elicit the evidence called for. Once the achievement levels, task models, and evidence models are established, the design proceeds by defining *task specifications* and *performance level descriptors (PLDs),* which contain all the preceding information in a form that lends itself to formulating the *test specifications.* These three components should be seen as components of an iterative process. As the name implies, task specifications are very specific descriptions of the tasks that will potentially comprise the assessment. It would be prudent to produce specifications for more tasks than can possibly be used in the assessment to allow for the possibility that some of them will not work out well. PLDs are (tentative) narratives of what students at each achievement levels can be said to know and are able to do.

A change to any of these components requires revisiting the other two; in practice, test specifications cannot be finalized without information about pragmatic constraints, such as budgets, testing time available, and so on. A requirement to shorten testing time would trigger changes to the test specifications, which in turn could trigger changes to the task specifications. Utmost care is needed in this process. Test specifications determine test-level attributes like reliability and decision consistency and need to be carefully thought through. An assessment that does classify students into achievement levels with sufficient consistency is a failure, no matter how soundly and carefully the achievement levels have been defined, since the uncertainty that will necessarily be attached to student-level and policy-level decisions based on such assessment will diminish its value.

This is an iterative process that aims at an optimal design, subject to relevant *pragmatic and psychometric constraints*. Note that among the psychometric constraints is incorporated the goal of achieving maximal discrimination in the region of the scale where the eventual cutscores are likely to be located. This is also an iterative process, ideally supplemented by field trials. Once the array of available tasks or task models is known and the constraints are agreed upon, a test *blueprint* can be formulated, which should be sufficiently detailed so that *preliminary cutscores* corresponding to the performance standards can be formulated. After the assessment is administered, it will be possible to evaluate whether the preliminary cutscores are well supported or need adjustment in light of the data that are available. At that juncture, the role of the standard setting panel is to accept the preliminary cutscores or to adjust them in the light of new information.

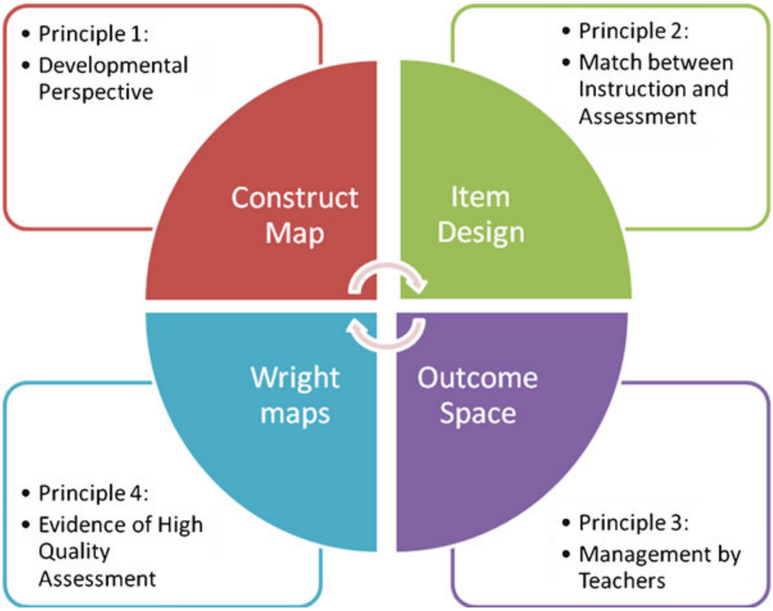


Fig. 3.19 The principles and building blocks of the BEAR Assessment System

The BEAR Assessment System[9]

As mentioned before, the *assessment structure* plays a key role in the in the study and the educational implementation of learning progressions. Although there are several alternative approaches that could be used to model, this section focuses in the BEAR Assessment System (BAS; Wilson 2005; Wilson and Sloane 2000), a measurement approach that will allow us to represent one of the various forms in which LPs could be conceived or measured.

The BEAR Assessment System is based on the idea that good assessment addresses the need for sound measurement by way of four principles: (1) a developmental perspective; (2) the match between instruction and assessment; (3) management by instructors to allow appropriate feedback, feedforward, and follow-up; and (4) generation of quality evidence. These four principles, with the four building blocks that embody them, are shown in Fig. 3.19. They serve as the basis of a model that is rooted in our knowledge of cognition and learning in each domain and that supports the alignment of instruction, curriculum, and assessment—all aspects recommended by the NRC (2001) as important components of educational assessment.

[9] The following section has been adapted from Wilson 2009.

Principle 1: A Developmental Perspective

A "developmental perspective" on student learning highlights two crucial ideas: (a) the need to characterize the evolution of learners over time and (b) the need for assessments that are "tailored" to the characteristics of different learning theories and learning domains.

The first element, portraying the evolution of learners over time, emphasizes the definition of relevant constructs based on the development of student mastery of particular concepts and skills over time, as opposed to making a single measurement at some final or supposedly significant point of time. Additionally, it promotes assessments based on "psychologically plausible" pathways of increasing proficiency, as opposed to attempt to assess contents based on logical approaches to the structure of disciplinary knowledge.

Much of the strength of the BEAR Assessment System is related to the second element, the emphasis on providing tools to model *many different kinds of learning theories and learning domains*, which avoids the "one-size-fits-all" development assessment approach that has rarely satisfied educational needs. What is to be measured and how it is to be valued in each BEAR assessment application is drawn from the expertise and learning theories of the teachers, curriculum developers, and assessment developers involved in the process of creating the assessments.

The developmental perspective assumes that student performance on a given learning progression can be traced over the course of instruction, facilitating a more developmental perspective on student learning. Assessing the growth of students' understanding of particular concepts and skills requires a model of how student learning develops over a certain period of (instructional) time; this growth perspective helps one to move away from "one shot" testing situations and cross-sectional approaches to defining student performance, toward an approach that focuses on the process of learning and on an individual's progress through that process. Clear definitions of what students are expected to learn and a theoretical framework of how that learning is expected to unfold as the student progresses through the instructional material (i.e., in terms of learning performances) are necessary to establish the construct validity of an assessment system.

Building Block 1: Construct Maps

Construct maps (Wilson 2005) embody this first of the four principles: a developmental perspective on assessing student achievement and growth. A construct map is a well-thought-out and researched ordering of qualitatively different levels of performance focusing on one characteristic that organizes clear definitions of the expected student progress. Thus, a construct map defines what is to be measured or assessed in terms general enough to be interpretable within a curriculum, and potentially across curricula, but specific enough to guide the development of the other components. When instructional practices are linked to the construct map, then the construct map also indicates the aims of the teaching.

Construct maps are derived in part from research into the underlying cognitive structure of the domain and in part from professional judgments about what constitutes higher and lower levels of performance or competence, but are also informed by empirical research into how students respond to instruction or perform in practice (NRC 2001).

Construct maps are one model of how assessments can be integrated with instruction and accountability. They provide a way for large-scale assessments to be linked in a principled way to what students are learning in classrooms, while having the potential at least to remain independent of the content of a specific curriculum.

The idea of using construct maps as the basis for assessments offers the possibility of gaining significant *efficiency* in assessment: Although each new curriculum prides itself on bringing something new to the subject matter, the truth is that most curricula are composed of a common stock of content. And, as the influence of national and state standards increases, this will become truer and make them easier to codify. Thus, we might expect innovative curricula to have one, or perhaps even two, variables that do not overlap with typical curricula, but the rest will form a fairly stable set of variables that will be common across many curricula.

Principle 2: Match Between Instruction and Assessment

The main motivation for the progress variables so far developed is that they serve as a framework for the assessments and a method for making measurement possible. However, this second principle makes clear that the framework for the assessments and the framework for the curriculum and instruction must be one and the same. This emphasis is consistent with research in the design of learning environments, which suggests that instructional settings should coordinate their focus on the learner (incorporated in Principle 1) with both knowledge-centered and assessment-centered environments (NRC 2000).

Building Block 2: The Items Design

The items design process governs the coordination between classroom instruction and assessment. The critical element to ensure this in the BEAR Assessment System is that each assessment task and typical student response is matched to particular levels of proficiency within at least one construct map.

When using this assessment system within a curriculum, a particularly effective mode of assessment is what is called *embedded assessment*. This means that opportunities to assess student progress and performance are integrated into the instructional materials and are (from the student's point of view) virtually indistinguishable from the day-to-day classroom activities.

It is useful to think of the metaphor of a stream of instructional activity and student learning, with the teacher dipping into the stream of learning from time to time to evaluate student progress and performance. In this model or metaphor, assessment then becomes part of the teaching and learning process, and we can think of it as being assessment for learning (AfL; Black et al. 2003).

If assessment is also a learning event, then it does not take time away from instruction unnecessarily, and the number of assessment tasks can be more readily increased so as to improve the reliability of the results (Linn and Baker 1996). Nevertheless, for assessment to become fully and meaningfully embedded in the teaching and learning process, the assessment must be linked to the curriculum and not be seen as curriculum-independent as is the rhetoric for traditional norm-referenced tests (Wolf and Reardon 1996).

Principle 3: Management by Teachers

For information from the assessment tasks and the BEAR analysis to be useful to instructors and students, it must be couched in terms that are directly related to the instructional goals behind the progress variables. Open-ended tasks, if used, must all be scorable—quickly, readily, and reliably.

Building Block 3: The Outcome Space

The outcome space is the set of categorical outcomes into which student performances are categorized, for all the items associated with a particular progress variable. In practice, these are presented as scoring guides for student responses to assessment tasks, which are meant to help make the performance criteria for the assessments clear and explicit (or "transparent and open" to use Glaser's (1963) terms)—not only to the teachers but also to the students and parents, administrators, or other "consumers" of assessment results. In fact, we strongly recommend to teachers that they share the scoring guides with administrators, parents, and students, as a way of helping them understand what types of cognitive performance are expected and to model the desired processes.

Scoring guides are the primary means by which the essential element of teacher professional judgment is implemented in the BEAR Assessment System. These are supplemented by "exemplars" of student work at every scoring level for each task and variable combination, and "blueprints," which provide the teachers with a layout indicating opportune times in the curriculum to assess the students on the different variables.

Principle 4: Evidence of High Quality Assessment

Technical issues of reliability and validity, fairness, consistency, and bias can quickly sink any attempt to measure along a progress variable, as described above, or even to develop a reasonable framework that can be supported by evidence. To ensure comparability of results across time and context, procedures are needed to (a) examine the coherence of information gathered using different formats, (b) map student performances onto the progress variables, (c) describe the structural elements of the accountability system—tasks and raters—in terms of the achievement variables, and (d) establish uniform levels of system functioning, in terms of quality control indices such as reliability.

Building Block 4: Wright Maps

Wright maps represent this principle of evidence of high quality. Wright maps are graphical and empirical representations of a construct map, showing how it unfolds or evolves in terms of increasingly sophisticated student performances.

They are derived from empirical analyses of student data on sets of assessment tasks. They show on an ordering of these assessment tasks from relatively easy tasks to more difficult ones. A key feature of these maps is that both students and tasks can be located on the same scale, giving student proficiency the possibility of substantive interpretation, in terms of what the student knows and can do and where the student is having difficulty. The maps can be used to interpret the progress of one particular student or the pattern of achievement of groups of students ranging from classes to nations.

Wright maps can be very useful in large-scale assessments, providing information that is not readily available through numerical score averages and other traditional summary information—they are used extensively, for example, in reporting on the PISA assessments (OECD 2005). Moreover, Wright maps can be seamlessly interpreted as representations of learning progressions, quickly mapping the statistical results back to the initial construct, providing the necessary evidence to explore questions about the structure of the learning progression, serving as the basis for improved versions of the original constructs.

References[*]

Abidi, S. S. R., Chong, Y., & Abidi, S. R. (2001). *Patient empowerment via 'pushed' delivery of customized healthcare educational content over the Internet*. Paper presented at the 10th World Congress on Medical Informatics, London.

Ackerman, T., Zhang, W., Henson, R., & Templin, J. (2006, April). *Evaluating a third grade science benchmark test using a skills assessment model: Q-matrix evaluation*. Paper presented at

[*] Note that this list also includes the references for the Annex

the annual meeting of the National Council on Measurement in Education (NCME), San Francisco

Adams, W. K., Reid, S., LeMaster, R., McKagan, S., Perkins, K., & Dubson, M. (2008). A study of educational simulations part 1—engagement and learning. *Journal of Interactive Learning Research, 19*(3), 397–419.

Aleinikov, A. G., Kackmeister, S., & Koenig, R. (Eds.). (2000). *101 Definitions: Creativity*. Midland: Alden B Dow Creativity Center Press.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment in Education*, *1*(5). Available from http://www.jtla.org

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2003). *A four-process architecture for assessment delivery, with connections to assessment design* (Vol. 616). Los Angeles: University of California Los Angeles Center for Research on Evaluations, Standards and Student Testing (CRESST).

American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (AERA, APA, NCME, 1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics, 118*(4), 1279–1333.

Ball, S. J. (1985). Participant observation with pupils. In R. Burgess (Ed.), *Strategies of educational research: Qualitative methods* (pp. 23–53). Lewes: Falmer.

Behrens, J. T., Frezzo D. C., Mislevy R. J., Kroopnick M., & Wise D. (2007). Structural, functional, and semiotic symmetries in simulation-based games and assessments. In E. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 59–80). New York: Earlbaum.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment, 2*(3). Available from http://www.jtla.org

Bejar, I. I., Braun, H., & Tannenbaum, R. (2007). A prospective, predictive and progressive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 1–30). Maple Grove: JAM Press.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9–16.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). New York: Springer.

Bennett, R. E., Goodman, M., Hessinger, J., Kahn, H., Ligget, J., & Marshall, G. (1999). Using multimedia in large-scale computer-based testing programs. *Computers in Human Behaviour, 15*(3–4), 283–294.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., & Rumble, M. (2009). *Developing 21st century skills and assessments*. White Paper from the Assessment and Learning of 21st Century Skills Project.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning*. London: Open University Press.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain*. New York/Toronto: Longmans, Green.

Bourque, M. L. (2009). *A history of NAEP achievement levels: Issues, implementation, and impact 1989–2009* (No. Paper Commissioned for the 20th Anniversary of the National Assessment

Governing Board 1988–2008). Washington, DC: NAGB. Downloaded from http://www.nagb.org/who-we-are/20-anniversary/bourque-achievement-levels-formatted.pdf

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 313–338). New York: Springer.

Braun, H., Bejar, I. I., & Williamson, D. M. (2006). Rule-based methods for automated scoring: Applications in a licensing context. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 83–122). Mahwah: Lawrence Erlbaum.

Brown, A. L., & Reeve, R. A. (1987). Bandwidths of competence: The role of supportive contexts in learning and development. In L. S. Liben (Ed.), *Development and learning: Conflict or congruence?* (pp. 173–223). Hillsdale: Erlbaum.

Brown, N. J. S., Furtak, E. M., Timms, M., Nagashima, S. O., & Wilson, M. (2010a). The evidence-based reasoning framework: Assessing scientific reasoning. *Educational Assessment, 15*(3–4), 123–141.

Brown, N. J. S., Nagashima, S. O., Fu, A., Timms, M., & Wilson, M. (2010b). A framework for analyzing scientific reasoning in assessments. *Educational Assessment, 15*(3–4), 142–174.

Brown, N., Wilson, M., Nagashima, S., Timms, M., Schneider, A., & Herman, J. (2008, March 24–28). *A model of scientific reasoning*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Brusilovsky, P., Sosnovsky, S., & Yudelson, M. (2006). Addictive links: The motivational value of adaptive link annotation in educational hypermedia. In V. Wade, H. Ashman, & B. Smyth (Eds.), *Adaptive hypermedia and adaptive Web-based systems, 4th International Conference, AH 2006*. Dublin: Springer.

Carnevale, A. P., Gainer, L. J., & Meltzer, A. S. (1990). *Workplace basics: The essential skills employers want*. San Francisco: Jossey-Bass.

Carpenter, T. P., & Lehrer, R. (1999). Teaching and learning mathematics with understanding. In E. Fennema & T. R. Romberg (Eds.), *Mathematics classrooms that promote understanding* (pp. 19–32). Mahwah: Lawrence Erlbaum Associates.

Case, R., & Griffin, S. (1990). Child cognitive development: The role of central conceptual structures in the development of scientific and social thought. In E. A. Hauert (Ed.), *Developmental psychology: Cognitive, perceptuo-motor, and neurological perspectives* (pp. 193–230). North-Holland: Elsevier.

Catley, K., Lehrer, R., & Reiser, B. (2005). *Tracing a prospective learning progression for developing understanding of evolution*. Paper Commissioned by the National Academies Committee on Test Design for K-12 Science Achievement. http://www7.nationalacademies.org/bota/Evolution.pdf

Center for Continuous Instructional Improvement (CCII). (2009). *Report of the CCII Panel on learning progressions in science* (CPRE Research Report). New York: Columbia University.

Center for Creative Learning. (2007). *Assessing creativity index*. Retrieved August 27, 2009, from http://www.creativelearning.com/Assess/index.htm

Chedrawy, Z., & Abidi, S. S. R. (2006). *An adaptive personalized recommendation strategy featuring context sensitive content adaptation.* Paper presented at the Adaptive Hypermedia and Adaptive Web-Based Systems, 4th International Conference, AH 2006, Dublin, Ireland.

Chen, Z.-L., & Raghavan, S. (2008). *Tutorials in operations research: State-of-the-art decision-making tools in the information-intensive age, personalization and recommender systems*. Paper presented at the INFORMS Annual Meeting. Retrieved from http://books.google.com/books?hl=en&lr=&id=4c6b1_emsyMC&oi=fnd&pg=PA55&dq=personalisation+online+entertainment+netflix&ots=haYV26Glyf&sig=kqjo5t1C1lNLlP3QG-R0iGQCG3o#v=onepage&q=&f=false

Claesgens, J., Scalise, K., Wilson, M., & Stacy, A. (2009). Mapping student understanding in chemistry: The perspectives of chemists. *Science Education, 93*(1), 56–85.

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences, 3*(9), 345–351.

Conlan, O., O'Keeffe, I., & Tallon, S. (2006). *Combining adaptive hypermedia techniques and ontology reasoning to produce Dynamic Personalized News Services*. Paper presented at the Adaptive Hypermedia and Adaptive Web-based Systems, Dublin, Ireland.

Crick, R. D. (2005). Being a Learner: A Virtue for the 21st Century. *British Journal of Educational Studies, 53*(3), 359–374.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Dagger, D., Wade, V., & Conlan, O. (2005). Personalisation for all: Making adaptive course composition easy. *Educational Technology & Society, 8*(3), 9–25.

Dahlgren, L. O. (1984). Outcomes of learning. In F. Martin, D. Hounsell, & N. Entwistle (Eds.), *The experience of learning*. Edinburgh: Scottish Academic Press.

DocenteMas. (2009). *The Chilean teacher evaluation system*. Retrieved from http://www.docentemas.cl/

Drasgow, F., Luecht, R., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Westport: Praeger Publishers.

Duncan, R. G., & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching, 46*(6), 606–609.

Frazier, E., Greiner, S., & Wethington, D. (Producer). (2004, August 14, 2009) *The use of biometrics in education technology assessment*. Retrieved from http://www.bsu.edu/web/elfrazier/TechnologyAssessment.htm

Frezzo, D. C., Behrens, J. T., & Mislevy, R. J. (2010). Design patterns for learning and assessment: Facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *Journal of Science Education and Technology, 19*(2), 105–114.

Frezzo, D. C., Behrens, J. T., Mislevy, R. J., West, P., & DiCerbo, K. E. (2009, April). Psychometric and evidentiary approaches to simulation assessment in Packet Tracer software. Paper presented at the Fifth International Conference on Networking and Services (ICNS), Valencia, Spain.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7*(4), 323–342.

Gellersen, H.-W. (1999). *Handheld and ubiquitous computing: First International Symposium.* Paper presented at the HUC '99, Karlsruhe, Germany.

Gifford, B. R. (2001). *Transformational instructional materials, settings and economics*. In The Case for the Distributed Learning Workshop, Minneapolis.

Giles, J. (2005). Wisdom of the crowd. Decision makers, wrestling with thorny choices, are tapping into the collective foresight of ordinary people. *Nature, 438*, 281.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *The American Psychologist, 18*, 519–521.

Graesser, A. C., Jackson, G. T., & McDaniel, B. (2007). AutoTutor holds conversations with learners that are responsive to their cognitive and emotional state. *Educational Technology, 47*, 19–22.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427–438.

Haladyna, T. M. (1994). Cognitive taxonomies. In T. M. Haladyna (Ed.), *Developing and validating multiple-choice test items* (pp. 104–110). Hillsdale: Lawrence Erlbaum Associates.

Hartley, D. (2009). Personalisation: The nostalgic revival of child-centred education? *Journal of Education Policy, 24*(4), 423–434.

Hattie, J. (2009, April 16). *Visibly learning from reports: The validity of score reports*. Paper presented at the annual meeting of the National Council on Measurement in Education (*NCME*), San Diego, CA.

Hawkins, D. T. (2007, November). Trends, tactics, and truth in the information industry: The fall 2007 ASIDIC meeting. *InformationToday*, p. 34.

Hayes, J. R. (1985). Three problems in teaching general skills. In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), *Thinking and learning skills: Research and open questions* (Vol. 2, pp. 391–406). Hillsdale: Erlbaum.

Henson, R., & Templin, J. (2008, March). *Implementation of standards setting for a geometry end-of-course exam*. Paper presented at the 2008 American Educational Research Association conference in New York, New York.

Hernández, J. A., Ochoa Ortiz, A., Andaverde, J., & Burlak, G. (2008). *Biometrics in online assessments: A study case in high school student.* Paper presented at the 8th International Conference on Electronics, Communications and Computers (conielecomp 2008), Puebla.

Hirsch, E. D. (2006, 26 April). Reading-comprehension skills? What are they really? *Education Week, 25*(33)*,* 57, 42.

Hopkins, D. (2004). *Assessment for personalised learning: The quiet revolution*. Paper presented at the Perspectives on Pupil Assessment, New Relationships: Teaching, Learning and Accountability, General Teaching Council Conference, London.

Howe, J. (2008, Winter). The wisdom of the crowd resides in how the crowd is used. *Nieman Reports, New Venues*, *62*(4), 47–50.

International Organization for Standardization. (2009). *International standards for business, government and society, JTC 1/SC 37—Biometrics.* http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_tc_browse.htm?commid=313770&development=on

Kanter, R. M. (1994). Collaborative advantage: The Art of alliances. *Harvard Business Review, 72*(4), 96–108.

Kelleher, K. (2006). Personalize it. *Wired Magazine, 14*(7), 1.

Kyllonen, P. C., Walters, A. M., & Kaufman, J. C. (2005). Noncognitive constructs and their assessment in graduate education: A review. *Educational Assessment, 10*(3), 143–184.

Lawton, D. L. (1970). *Social class, language and education*. London: Routledge and Kegan Paul.

Lesgold, A. (2009). *Better schools for the 21st century: What is needed and what will it take to get improvement*. Pittsburgh: University of Pittsburgh.

Levy, F., & Murnane, R. (2006, May 31). *How computerized work and globalization shape human skill demands*. Retrieved August 23, 2009, from http://web.mit.edu/flevy/www/computers_offshoring_and_skills.pdf

Linn, R. L., & Baker, E. L. (1996). Can performance- based student assessments be psychometrically sound? In J. B. Baron, & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (pp. 84–103). Chicago: University of Chicago Press.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694.

Lord, F. M. (1971). Tailored testing, an approximation of stochastic approximation. *Journal of the American Statistical Association, 66*, 707–711.

Margolis, M. J., & Clauser, B. E. (2006). A regression-based procedure for automated scoring of a complex medical performance assessment. In D. M. Williamson, I. J. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer based testing*. Mahwah: Lawrence Erlbaum Associates.

Martinez, M. (2002). What is personalized learning? Are we there yet? *E-Learning Developer's Journal*. E-Learning Guild (www.elarningguild.com). http://www.elearningguild.com/pdf/2/050702dss-h.pdf

Marton, F. (1981). Phenomenography—Describing conceptions of the world around us. *Instructional Science, 10*, 177–200.

Marton, F. (1983). Beyond individual differences. *Educational Psychology, 3*, 289–303.

Marton, F. (1986). Phenomenography—A research approach to investigating different understandings of reality. *Journal of Thought, 21*, 29–49.

Marton, F. (1988). Phenomenography—Exploring different conceptions of reality. In D. Fetterman (Ed.), *Qualitative approaches to evaluation in education* (pp. 176–205). New York: Praeger.

Marton, F., Hounsell, D., & Entwistle, N. (Eds.). (1984). *The experience of learning*. Edinburgh: Scottish Academic Press. Masters, G.N. & Wilson, M. (1997). *Developmental assessment*. Berkeley, CA: BEAR Research Report, University of California.

Masters G. (1982). *A rasch model for partial credit scoring.* Psychometrika 42(2), 149–174.

Masters, G.N. & Wilson, M. (1997). Developmental assessment. Berkeley, CA: BEAR Research Report, University of California.

Mayer, R. E. (1983). *Thinking, problem-solving and cognition*. New York: W H Freeman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education/Macmillan.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *The American Psychologist, 50*(9), 741–749.

Microsoft. (2009). *Microsoft Certification Program*. Retrieved from http://www.microsoft.com/learning/

Miliband, D. (2003). Opportunity for all, targeting disadvantage through personalised learning. *New Economy, 1070–3535/03/040224*(5), 224–229.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003a). *A brief introduction to evidence centred design* (Vol. RR-03–16). Princeton: Educational Testing Service.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003b). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Mislevy, R. J., Bejar, I. I., Bennett, R. E., Haertel, G. D., & Winters, F. I. (2008). Technology supports for assessment design. In B. McGaw, E. Baker, & P. Peterson (Eds.), *International encyclopedia of education* (3rd ed.). Oxford: Elsevier.

Mitchell, W. J. (1990). *The logic of architecture*. Cambridge: MIT Press.

National Research Council, Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. Washington, DC: National Academy Press.

National Research Council, Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

National Research Council, Wilson, M., & Bertenthal, M. (Eds.). (2006). *Systems for state science assessment. Committee on Test Design for K-12 Science Achievement*. Washington, DC: National Academy Press.

National Research Council, Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in Grades K-8*. Committee on Science Learning, Kindergarten through Eighth Grade. Washington, DC: National Academy Press.

Newell, A., Simon, H. A., & Shaw, J. C. (1958). Elements of a theory of human problem solving. *Psychological Review, 65*, 151–166.

Oberlander, J. (2006). *Adapting NLP to adaptive hypermedia.* Paper presented at the Adaptive Hypermedia and Adaptive Web-Based Systems, 4th International Conference, AH 2006, Dublin, Ireland.

OECD. (2005). *PISA 2003 Technical Report*. Paris: Organisation for Economic Co-operation and Development.

Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation, 13*(4), 4.

Parshall, C. G., Stewart, R., Ritter, J. (1996, April). *Innovations: Sound, graphics, and alternative response modes*. Paper presented at the National Council on Measurement in Education, New York.

Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. Van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 129–148). Norwell: Kluwer Academic Publisher.

Parshall, C. G., Spray, J., Kalohn, J., & Davey, T. (2002). *Issues in innovative item types practical considerations in computer-based testing* (pp. 70–91). New York: Springer.

Patton, M. Q. (1980). *Qualitative evaluation methods*. Beverly Hills: Sage.

Pellegrino, J., Jones, L., & Mitchell, K. (Eds.). (1999). *Grading the Nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press.

Perkins, D. (1998). What is understanding? In M. S. Wiske (Ed.), *Teaching for understanding: Linking research with practice*. San Francisco: Jossey-Bass Publishers.

Pirolli, P. (2007). *Information foraging theory: Adaptive interaction with information*. Oxford: Oxford University Press.

Popham, W. J. (1997). Consequential validity: Right concern—Wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–13.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14*, 58–93.

Reiser, R. A. (2002). A history of instructional design and technology. In R. A. Reiser & J. V. Dempsey (Eds.), *Trends and issues in instructional design and technology*. Upper Saddle River: Merrill/Prentice Hall.

Reiser, B., Krajcik, J., Moje, E., & Marx, R. (2003, March). *Design strategies for developing science instructional materials*. Paper presented at the National Association for Research in Science Teaching, Philadelphia, PA.

Robinson, K. (2009). *Out of our minds: Learning to be creative*. Chichester: Capstone.

Rosenbaum, P. R. (1988). Item Bundles. *Psychometrika, 53*, 349–359.

Rupp, A. A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*, 219–262.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.

Scalise, K. (2004). *A new approach to computer adaptive assessment with IRT construct-modeled item bundles (testlets): An application of the BEAR assessment system*. Paper presented at the 2004 International Meeting of the Psychometric Society, Pacific Grove.

Scalise, K. (submitted). Personalised learning taxonomy: Characteristics in three dimensions for ICT. *British Journal of Educational Technology*.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in E-Learning: A framework for constructing "Intermediate Constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4(6)* [online journal]. http://escholarship.bc.edu/jtla/vol4/6.

Scalise, K., & Wilson, M. (2006). Analysis and comparison of automated scoring approaches: Addressing evidence-based assessment principles. In D. M. Williamson, I. J. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer based testing*. Mahwah: Lawrence Erlbaum Associates.

Scalise, K., & Wilson, M. (2007). *Bundle models for computer adaptive testing in e-learning assessment*. Paper presented at the 2007 GMAC Conference on Computerized Adaptive Testing (Graduate Management Admission Council), Minneapolis, MN.

Schum, D. A. (1987). *Evidence and inference for the intelligence analyst*. Lanham: University Press of America.

Searle, J. (1969). *Speech acts*. Cambridge: Cambridge University Press.

Shute, V., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). *Melding the power of serious games and embedded assessment to monitor and foster learning, flow and grow melding the power of serious games*. New York: Routledge.

Shute, V., Maskduki, I., Donmez, O., Dennen, V. P., Kim, Y. J., & Jeong, A. C. (2010). Modeling, assessing, and supporting key competencies within game environments. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge*. New York: Springer. Smith, C., Wiser, M., Anderson, C. W., Krajcik, J. & Coppola, B. (2004). *Implications of research on children's learning for assessment: matter and atomic* *molecular theory*. Paper Commissioned by the National Academies Committee on Test Design for K-12 Science Achievement. Washington DC

Simon, H. A. (1980). Problem solving and education. In D. T. Tuma, & R. Reif, (Eds.), *Problem solving and education: Issues in teaching and research* (pp. 81–96). Hillsdale: Erlbaum.

Smith, C., Wiser, M., Anderson, C. W., Krajcik, J. & Coppola, B. (2004). *Implications of research on children's learning for assessment: matter and atomic molecular theory*. Paper Commissioned by the National Academies Committee on Test Design for K-12 Science Achievement. Washington DC.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on Children's learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*(1 & 2).

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758–765.

Templin, J., & Henson, R. A. (2008, March). *Understanding the impact of skill acquisition: relating diagnostic assessments to measureable outcomes*. Paper presented at the 2008 American Educational Research Association conference in New York, New York.

Treffinger, D. J. (1996). *Creativity, creative thinking, and critical thinking: In search of definitions*. Sarasota: Center for Creative Learning.

Valsiner, J., & Veer, R. V. D. (2000). *The social mind*. Cambridge: Cambridge University Press.

Van der Linden, W. J., & Glas, C. A. W. (2007). Statistical aspects of adaptive testing. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 801–838). New York: Elsevier.

Wainer, H., & Dorans, N. J. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.

Wainer, H., Brown, L., Bradlow, E., Wang, X., Skorupski, W. P., & Boulet, J. (2006). An application of testlet response theory in the scoring of a complex certification exam. In D. M. Williamson, I. J. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer based testing*. Mahwah: Lawrence Erlbaum Associates.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126–149.

Weekley, J. A., & Ployhart, R. E. (2006). *Situational judgment tests: Theory, measurement, and application*. Mahwah: Lawrence Erlbaum Associates.

Weiss, D. J. (Ed.). (2007). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. Available at http://www.psych.umn.edu/psylabs/catcentral/

Wiley, D. (2008). *Lying about personalized learning, iterating toward openness*. Retrieved from http://opencontent.org/blog/archives/655

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Mahwah: Lawrence Erlbaum Associates.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah: Lawrence Erlbaum Associates.

Wilson, M. (Ed.). (2004). *Towards coherence between classroom assessment and accountability*. Chicago: Chicago University Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah: Lawrence Erlbaum Associates.

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika, 60*(2), 181–198.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*, 181–208.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(6), 716–730.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*, 19–38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163–183.

Wolf, D. P., & Reardon, S. F. (1996). Access to excellence through new forms of student assessment. In D. P. Wolf, & J. B. Baron (Eds.), *Performance-based student assessment: Challenges and possibilities. Ninety-fifth yearbook of the national society for the study of education, part I*. Chicago: University of Chicago Press.

Zechner, K., Higgins, D., Xiaoming, X., & Williamson, D. (2009). Automatic scoring of non-native spontaneous speech in test of spoken English. *Speech Communication, 51*, 883–895.

Chapter 4
Technological Issues for Computer-Based Assessment

Benő Csapó, John Ainley, Randy E. Bennett, Thibaud Latour, and Nancy Law

Abstract This chapter reviews the contribution of new information-communication technologies to the advancement of educational assessment. Improvements can be described in terms of precision in detecting the actual values of the observed variables, efficiency in collecting and processing information, and speed and frequency of feedback given to the participants and stakeholders. The chapter reviews previous research and development in two ways, describing the main tendencies in four continents (Asia, Australia, Europe and the US) as well as summarizing research on how technology advances assessment in certain crucial dimensions (assessment of established constructs, extension of assessment domains, assessment of new constructs and in dynamic situations). As there is a great variety of applications of assessment in education, each one requiring different technological solutions, the chapter classifies assessment domains, purposes and contexts and identifies the technological needs and solutions for each. The chapter reviews the contribution of technology to the advancement of the entire educational evaluation process, from authoring and automatic generation and storage of items, through delivery methods (Internet-based, local server, removable media, mini-computer labs) to forms of task presentation made possible with technology for response capture, scoring and automated feedback and reporting. Finally, the chapter identifies areas for which further

B. Csapó (✉)
Institute of Education, University of Szeged,
e-mail: csapo@edpsy.u-szeged.hu

J. Ainley
Australian Council for Educational Research

R.E. Bennett
Educational Testing Service, Princeton

T. Latour
Henri Tudor Public Research Centre, Luxembourg

N. Law
Faculty of Education, University of Hong Kong

P. Griffin et al. (eds.), *Assessment and Teaching of 21st Century Skills*,
DOI 10.1007/978-94-007-2324-5_4, © Springer Science+Business Media B.V. 2012

research and development is needed (migration strategies, security, availability, accessibility, comparability, framework and instrument compliance) and lists themes for research projects feasible for inclusion in the *Assessment and Teaching of Twenty-first Century Skills project.*

Information–communication technology (ICT) offers so many outstanding possibilities for teaching and learning that its application has been growing steadily in every segment of education. Within the general trends of the use of ICT in education, technology-based assessment (TBA) represents a rapidly increasing share. Several traditional assessment processes can be carried out more efficiently by means of computers. In addition, technology offers new assessment methods that cannot be otherwise realized. There is no doubt that TBA will replace paper-based testing in most of the traditional assessment scenarios, and technology will further extend the territories of assessment in education as it provides frequent and precise feedback for participants in learning and teaching that cannot be achieved by any other means.

At the same time, large-scale implementation of TBA still faces several technological challenges that need further research and a lot of experimentation in real educational settings. The basic technological solutions are already available, but their application in everyday educational practice, especially their integration into educationally optimized, consistent systems, requires further developmental work.

A variety of technological means operate in schools, and the diversity, compatibility, connectivity and co-working of those means require further considerations. Each new technological innovation finds its way to schools but not always in a systematic way. Thus, the possibilities of technology-driven modernization of education—when the intent to apply emerging technological tools motivates changes—are limited. In this chapter, another approach is taken in which the actual and conceivable future problems of educational development are considered and the available technological means are evaluated according to their potential to contribute in solving them.

Technology may significantly advance educational assessment along a number of dimensions. It improves the precision of detecting the actual values of the observed variables and the efficiency of collecting and processing information; it enables the sophisticated analysis of the available data, supports decision-making and provides rapid feedback for participants and stakeholders. Technology helps to detect and record the psychomotor, cognitive and affective characteristics of students and the social contexts of teaching and learning processes alike. When we deal with technological issues in educational assessment, we limit our analyses of the human side of the human–technology interaction. Although technological problems in a narrow sense, like the parameters of the available instruments—e.g. processor speed, screen resolution, connection bandwidth—are crucial in educational application, these questions play a secondary role in our study. In this chapter, we mostly use the more general term *technology-based assessment*, meaning that there are several technical tools beyond the most commonly used computers. Nevertheless, we are aware that in the foreseeable future, computers will continue to play a dominant role.

The entire project focuses on the twenty-first-century skills; however, when dealing with technological issues, we have to consider a broader perspective. In this chapter, our position concerning twenty-first-century skills is that we are not dealing exclusively with them because:

- They are not yet identified with sufficient precision and accuracy that their definition could orient the work concerning technological issues.
- We assume that they are based on certain basic skills and 'more traditional' sub-skills and technology should serve the assessment of those components as well.
- In the real educational context, assessment of twenty-first-century skills is not expected to be separated from the assessment of other components of students' knowledge and skills; therefore, the application of technology needs to cover a broader spectrum.
- Several of the technologies used today for the assessment of students' knowledge may be developed and adapted for the specific needs of the assessment of twenty-first-century skills.
- There are skills that are obviously related to the modern, digital world, and technology offers excellent means to assess them; so we deal with these specific issues whenever appropriate throughout the chapter (e.g. dynamic problem-solving, complex problem-solving in technology-rich environment, working in groups whose members are connected by ICT).

Different assessment scenarios require different technological conditions, so one single solution cannot optimally serve every possible assessment need. Teaching and learning in a modern society extend well beyond formal schooling, and even in traditional educational settings, there are diverse forms of assessment, which require technologies adapted to the actual needs. Different technological problems have to be solved when computers are used to administer high-stakes, large-scale, nationally or regionally representative assessments under standardized conditions, as well as low-stakes, formative, diagnostic assessment in a classroom environment under diverse school conditions. Therefore, we provide an overview of the most common assessment types and identify their particular technological features.

Innovative assessment instruments raise several methodological questions, and it requires further analysis on how data collection with the new instruments can satisfy the basic assumptions of psychometrics and on how they fit into the models of classical or modern test theories. This chapter, in general, does not deal with methodological questions. There is one methodological issue that should be considered from a technological point of view, however, and this is validity. Different validity issues may arise when TBA is applied to replace traditional paper-based assessment and when skills related to the digital world are assessed.

In this chapter, technological issues of assessment are considered in a broader sense. Hence, beyond reviewing the novel data collection possibilities, we deal with the questions of how technology may serve the entire educational evaluation process, including item generation, automated scoring, data processing, information flow, feedback and supporting decision-making.

Conceptualizing Technology-Based Assessment

Diversity of Assessment Domains, Purposes and Contexts

Assessment occurs in diverse domains for a multiplicity of purposes and in a variety of contexts for those being assessed. Those domains, purposes and contexts are important to identify because they can have implications for the ways that technology might be employed to improve testing and for the issues associated with achieving that improvement.

Assessment Domains

The relationship between domain or construct definition and technology is critical because it influences the role that technology can play in assessment. Below, we distinguish five general situations, each of which poses different implications for the role that technology might play in assessment.

The first of these is characterized by domains in which practitioners interact with the new technology primarily using specialized tools, if they use technology tools at all. In mathematics, such tools as symbol manipulators, graphing calculators and spreadsheets are frequently used—but typically only for certain purposes. For many mathematical problem-solving purposes, paper and pencil remains the most natural and fastest way to address a problem, and most students and practitioners use that medium a significant proportion of the time. It would be relatively rare for a student to use technology tools exclusively for mathematical problem-solving. For domains in this category, testing with technology needs either to be restricted to those problem-solving purposes for which technology is typically used or be implemented in such a way as not to compromise the measurement of those types of problem-solving in which technology is not usually employed (Bennett et al. 2008).

The second situation is characterized by those domains in which, depending upon the preferences of the individual, technology may be used exclusively or not at all. The domain of writing offers the clearest example. Not only do many practitioners and students routinely write on computer, many individuals virtually do their entire academic and workplace writing on computer. Because of the facility provided by the computer, they may write better and faster in that mode than they could on paper. Other individuals still write exclusively on paper; for these students and practitioners, the computer is an impediment because they haven't learned how to use it in composition. For domains of this second category, testing with technology can take three directions, depending upon the information needs of test users: (1) testing all students in the traditional mode to determine how effectively they perform in that mode, (2) testing all students with technology to determine how proficient they are in applying technology in that domain or (3) testing students in the mode in which they customarily work (Horkay et al. 2006).

The third situation is defined by those domains in which technology is so central that removing it would render it meaningless. The domain of computer programming would be an example; that domain cannot be effectively taught or practised without using computers. For domains of this category, proficiency cannot be effectively assessed unless all individuals are tested through technology (Bennett et al. 2007).

The fourth situation relates to assessing whether someone is capable of achieving a higher level of performance with the appropriate use of general or domain-specific technology tools than would be possible without them. It differs from the third situation in that the task may be performed without the use of tools, but only by those who have a high-level mastery of the domain and often in rather cumbersome ways. Here the tools are those that are generally referred to as cognitive tools, such as simulations and modelling tools (Mellar et al. 1994; Feurzeig and Roberts 1999), geographic information systems (Kerski 2003; Longley 2005) and visualization tools (Pea 2002).

The fifth situation relates to the use of technology to support collaboration and knowledge building. It is commonly acknowledged that knowledge creation is a social phenomenon achieved through social interactions, even if no direct collaboration is involved (Popper 1972). There are various projects on technology-supported learning through collaborative inquiry in which technology plays an important role in the provision of cognitive and metacognitive guidance (e.g. in the WISE project, see Linn and Hsi 1999). In some cases, the technology plays a pivotal role in supporting the socio-metacognitive dynamics that are found to be critical to productive knowledge building (Scardamalia and Bereiter 2003), since knowledge building is not something that happens naturally but rather has to be an intentional activity at the community level (Scardamalia 2002).

Thus, how a domain is practised, taught and learned influences how it should be assessed because misalignment of assessment and practice methods can compromise the meaning of assessment results. Also, it is important to note that over time, domain definitions change because the ways that they are practised and taught change, a result in part of the emergence of new technology tools suited to these domains. Domains that today are characterized by the use of technology for specialized purposes only may tomorrow see a significant proportion of individuals employing technology as their only means of practice. As tools advance, technology could become central to the definition of those domains too.

Of the five domains of technology use described above, the third, fourth and fifth domains pose the greatest challenge to assessment, and yet it is exactly these that are most important to include in the assessment of twenty-first-century skills since 'the real promise of technology in education lies in its potential to facilitate fundamental, qualitative changes in the nature of teaching and learning' (Panel on Educational Technology of the President's Committee of Advisors on Science and Technology 1997, p. 33).

Assessment Purposes

Here, we distinguish four general purposes for assessment, deriving from the two-way classification of assessment 'object' and assessment 'type'. The object of assessment may be the student, or it may be a programme or institution. Tests administered for purposes of drawing conclusions about programs or institutions have traditionally been termed ‘program evaluation’. Tests given for drawing conclusions about individuals have often been called ‘assessment’.

For either programme evaluation or assessment, two types can be identified: formative versus summative (Bloom 1969; Scriven 1967). Formative evaluation centres upon providing information for purposes of programme improvement, whereas summative evaluation focuses on judging the overall value of a programme. Similarly, formative assessment is intended to provide information of use to the teacher or student in modifying instruction, whereas summative assessment centres upon documenting what a student (or group of students) knows and can do.

Assessment Contexts

The term assessment context generally refers to the stakes that are associated with decisions based on test performance. The highest stakes are associated with those decisions that are serious in terms of their impact on individuals, programmes or institutions and that are not easily reversible. The lowest stakes are connected to decisions that are likely to have less impact and that are easily reversible. While summative measures have typically been taken as high stakes and formative types as low stakes, such blanket classifications may not always hold, if only because a single test may have different meanings for different constituencies. The US National Assessment of Educational Progress (NAEP) is one example of a summative test in which performance has low stakes for students, as no individual scores are computed, but high stakes for policymakers, whose efforts are publicly ranked. A similar situation obtains for summative tests administered under the US *No Child Left Behind* act, where the results may be of no consequence to students, while they have major consequences for individual teachers, administrators and schools. On the other hand, a formative assessment may involve low stakes for the school but considerable stakes for a student if the assessment directs that student towards developing one skill to the expense of another one more critical to that student’s short-term success (e.g. in preparing for an upcoming musical audition).

The above definition of context can be adequate if the assessment domain is well understood and assessment methods are well developed. If the domains of assessment and/or assessment methods (such as using digital technology to mediate the delivery of the assessment) are new, however, rather different considerations of design and method are called for. To measure more complex understanding and skills, and to integrate the use of technology into the assessment process so as to reflect such new learning outcomes, requires innovation in assessment (Quellmalz and Haertel 2004). In such situations, new assessment instruments probably have to be developed or invented, and it is apparent that both the validity and reliability can only be refined and established over a period of time, even if the new assessment domain is well defined. For assessing twenty-first-century skills, this kind of contextual challenge is even greater, since what constitute the skills to be assessed are, in themselves, a subject of debate. How innovative assessment can provide formative feedback on curriculum innovation and vice versa is another related challenge.

Using Technology to Improve Assessment

Technology can be used to improve assessment in at least two major ways: by changing the business of assessment and by changing the substance of assessment itself (Bennett 2001). The business of assessment means the core processes that define the enterprise. Technology can help make these core processes more efficient. Examples can be found in:

- Developing tests, making the questions easier to generate automatically or semi-automatically, to share, review and revise (e.g. Bejar et al. 2003)
- Delivering tests, obviating the need for printing, warehousing and shipping paper instruments
- Presenting dynamic stimuli, like audio, video and animation, making obsolete the need for specialized equipment currently being used in some testing programmes for assessing such constructs as speech and listening (e.g. audio cassette recorders, VCRs) (Bennett et al. 1999)
- Scoring constructed responses on screen, allowing marking quality to be monitored in real time and potentially eliminating the need to gather examiners together (Zhang et al. 2003)
- Scoring some types of constructed responses automatically, reducing the need for human reading (Williamson et al. 2006b)
- Distributing test results, cutting the costs of printing and mailing reports

Changing the substance of assessment involves using technology to change the nature of what is tested, or learned, in ways not practical with traditional assessment approaches or with technology-based duplications of those approaches (as by using a computer to record an examinee's speech in the same way as a tape recorder). An example would be asking students to experiment with and draw conclusions from an interactive simulation of a scientific phenomenon they could otherwise not experience and then using features of their problem-solving processes to make judgements about those students (e.g. Bennett et al. 2007). A second example would be in structuring the test design so that students learn in the process of taking the assessment by virtue of the way in which the assessment responds to student actions.

The use of technology in assessment may also play a crucial role in informing curriculum reform and pedagogical innovation, particularly in areas of specific domains in which technology has become crucial to the learning. For example, the Hong Kong SAR government commissioned a study to conduct online performance assessment of students' information literacy skills as part of the evaluation of the effectiveness of its IT in education strategies (Law et al. 2007). In Hong Kong, an important premise for the massive investments to integrate IT in teaching and learning is to foster the development of information literacy skills in students so that they can become more effective lifelong learners and can accomplish the learning in the designated curriculum more effectively. The study assessed students' ability to search for and evaluate information, and to communicate and collaborate with distributed peers in the context of authentic problem-solving through an online platform. The study found that while a large majority of the assessed students were able to demonstrate basic technical operational skills, their ability to demonstrate higher levels of cognitive functioning, such as evaluation and integration of information, was rather weak. This led to new initiatives in the Third IT in Education Strategy (EDB 2007) to develop curriculum resources and self-access assessment tools on information literacy. This is an example in which assessment has been used formatively to inform and improve on education policy initiatives.

The ways that technology might be used to improve assessment, while addressing the issues encountered, all depend on the domain, purpose and context of assessment. For example, fewer issues might be encountered when implementing formative assessments in low-stakes contexts targeted at domains where technology is central to the domain definition than for summative assessments in high-stakes contexts where technology is typically used only for certain types of problem-solving.

Review of Previous Research and Development

Research and development is reviewed here from two different viewpoints. On the one hand, a large number of research projects have been dealing with the application of technology to assessment. The devices applied in the experiments may range from the most common, widely available computers to emerging cutting-edge technologies. For research purposes, newly developed expensive instruments may be used, and specially trained teachers may participate; therefore, these experiments are often at small scale, carried out in a laboratory context or involving only a few classes or schools.

On the other hand, there are efforts for system-wide implementation of TBA either to extend, improve or replace the already existing assessment systems or to create entirely new systems. These implementation processes usually involve nationally representative samples from less than a thousand up to several thousand students. Large international programmes aim as well at using technologies for assessment, with the intention of both replacing paper-based assessment by TBA and introducing innovative domains and contexts that cannot be assessed by traditional testing methods. In large-scale implementation efforts, the general educational contexts (school infrastructure) are usually given, and either the existing equipment is used as it is, or new equipment is installed for assessment purposes. Logistics in these cases plays a crucial role; furthermore, a number of financial and organizational aspects that influence the choice of the applicable technology have to be considered.

Research on Using Technology for Assessment

ICT has already begun to alter educational assessment and has potential to change it further. One aspect of this process has been the more effective and efficient delivery of traditional assessments (Bridgeman 2009). A second has been the use of ICT to expand and enrich assessment tools so that assessments better reflect the intended domains and include more authentic tasks (Pellegrino et al. 2004). A third aspect has been the assessment of constructs that either have been difficult to assess or have emerged as part of the information age (Kelley and Haber 2006). A fourth has been the use of ICT to investigate the dynamic interactions between student and assessment material.

Published research literature on technology and computer-based assessment predominantly reflects research comparing the results of paper-based and computer-based assessment of the same construct. This literature seeks to identify the extent to which these two broad modalities provide congruent measures. Some of that literature draws attention to the importance of technological issues (within computer-based assessments) on measurement. There is somewhat less literature concerned with the properties of assessments that deliberately seek to extend the construct being assessed by making use of the possibilities that arise from computer-based assessment. An even more recent development has been the use of computer-based methods to assess new constructs: those linked to information technology, those using computer-based methods to assess constructs that have been previously hard to measure or those based on the analysis of dynamic interactions. The research literature on these developments is limited at this stage but will grow as the applications proliferate.

Assessment of Established Constructs

One important issue in the efficient delivery of assessments has been the equivalence of the scores on computer-administered assessments to those on the corresponding paper-based tests. The conclusion of two meta-analyses of studies of computer-based assessments of reading and mathematics among school students is that overall, the mode of delivery does not affect scores greatly (Wang et al. 2007, 2008). This generalization appears to hold for small-scale studies of abilities (Singleton 2001), large-scale assessments of abilities (Csapó et al. 2009) and large-scale assessments of achievement (Poggio et al. 2004). The same generalization appears to have been found true in studies conducted in higher education. Despite this overall result, there do appear to be some differences in scores associated with some types of questions and some aspects of the ways that students approach tasks (Johnson and Green 2006). In particular, there appears to be an effect of computer familiarity on performance in writing tasks (Horkay et al. 2006).

Computer-based assessment, in combination with modern measurement theory, has given impetus to expanding the possibility of computer adaptive testing

(Wainer 2000; Eggen and Straetmans 2009). Computer adaptive testing student performance on items is dynamic, meaning that subsequent items are selected from an item bank at a more appropriate difficulty for that student, providing more time-efficient and accurate assessments of proficiency. Adaptive tests can provide more evenly spread precision across the performance range, are shorter for each person assessed and maintain a higher level of precision overall than a fixed-form test (Weiss and Kingsbury 2004). However, they are dependent on building and calibrating an extensive item bank.

There have been a number of studies of variations within a given overall delivery mode that influence a student's experience of an assessment. There is wide acceptance that it is imperative for all students to experience the tasks or items presented in a computer-based assessment in an identical manner. Uniformity of presentation is assured when students are given the assessment tasks or items in a test booklet. However, there is some evidence that computer-based assessment can affect student performance because of variations in presentation not relevant to the construct being assessed (Bridgeman et al. 2003; McDonald 2002). Bridgeman et al. (2003) point out the influence of variations in screen size, screen resolution and display rate on performance on computer-based assessments. These are issues in computer-based assessments that do not normally arise in pen-and-paper assessments. Thompson and Weiss (2009) argue that the possibilities of variations in the assessment experience are a particular issue for Internet- or Web-based delivery of assessments, important considerations for the design of assessment delivery systems. Large-scale assessments using ICT face the problem of providing a uniform testing environment when school computing facilities can vary considerably.

Extending Assessment Domains

One of the issues confronting assessment has been that what could be assessed by paper-based methods represents a narrower conception of the domain than one would ideally wish for. The practice of assessment has been limited by what could be presented in a printed form and answered by students in writing. Attempts to provide assessments of broader aspects of expertise have been limited by the need to be consistent and, in the case of large-scale studies, a capacity to process rich answers. In many cases, these pressures have resulted in the use of closed-response formats (such as multiple choice) rather than constructed response formats in which students write a short or extended answer.

ICT can be used to present richer stimulus material (e.g. video or richer graphics), to provide for students to interact with the assessment material and to develop products that are saved for subsequent assessment by raters. In the *Programme for International Student Assessment* (PISA) 2006, a computer-based assessment of science (CBAS) was developed for a field trial in 13 countries and implemented as a main survey in three countries (OECD 2009, 2010). It was then adopted as part of the main study in three countries. CBAS was intended to assess aspects of science that could not be assessed in paper-based formats, so it involved an extension of the implemented assessment domain while not attempting to cover the whole of the domain. It was based on providing rich stimulus material linked to conventional test item formats. The design for the field trial included a rotated design that had half of the students doing a paper-based test first, followed by a computer test and the other half doing the tests in the opposite order. In the field trial, the correlation between the paper-based and computer-based items was 0.90, but it was also found that a two-dimensional model (dimensions corresponding to the paper- and computer-based assessment items) was a better fit than a one-dimensional model (Martin et al. 2009). This suggests that the dimension of science knowledge and understanding represented in the CBAS items was related to, but somewhat different from, the dimension represented in the paper-based items. Halldórsson et al. (2009) showed that, in the main PISA survey in Iceland, boys performed relatively better than girls but that this difference was not associated with differences in computer familiarity, motivation or effort. Rather, it did appear to be associated with the lower reading load on the computer-based assessment. In other words, the difference was not a result of the mode of delivery as such but of a feature that was associated with the delivery mode: the amount of text to be read. At present, reading is modified on the computer because of restrictions of screen size and the need to scroll to see what would be directly visible in a paper form. This limitation of the electronic form is likely to be removed as *e-book* and other developments are advanced.

Assessing New Constructs

A third focus on research on computer-based assessment is on assessing new constructs. Some of these relate directly to skills either associated with information technology or changed in nature as a result of its introduction. An example is ‘problem solving in rich technology environments’ (Bennett et al. 2010). Bennett et al. (2010) measured this construct in a nationally (USA) representative sample of grade 8 students. The assessment was based on two extended scenarios set in the context of scientific investigation: one involving a search and the other, a simulation. The *Organization for Economic Co-operation and Development* (OECD) *Programme for International Assessment of Adult Competencies* (PIAAC) includes ‘problem solving in technology-rich environments’ as one of the capabilities that it assesses among adults (OECD 2008b). This refers to the cognitive skills required in the information age, focussed on solving problems using multiple sources of information on a laptop computer. The problems are intended to involve accessing, evaluating, retrieving and processing information and incorporate technological and cognitive demands.

Wirth and Klieme (2003) investigated analytical and dynamic aspects of problem-solving. Analytical abilities were those needed to structure, represent and integrate information, whereas dynamic problem-solving involved the ability to adapt to a changing environment by processing feedback information (and included aspects of self-regulated learning). As a German national option in PISA 2000, the analytical and dynamic problem-solving competencies of 15-year-old students were tested using paper-and-pencil tests as well as computer-based assessments. Wirth and Klieme reported that analytical aspects of problem-solving competence were strongly correlated with reasoning, while dynamic problem-solving reflected a dimension of self-regulated exploration and control that could be identified in computer-simulated domains.

Another example of computer-based assessment involves using new technology to assess more enduring constructs, such as teamwork (Kyllonen 2009). *Situational Judgment Tests* (SJTs) involve presenting a scenario (incorporating audio or video) involving a problem and asking the student the best way to solve it. A meta-analysis of the results of several studies of SJTs of teamwork concluded that they involve both cognitive ability and personality attributes and predict real-world outcomes (McDaniel et al. 2007). Kyllonen argues that SJTs provide a powerful basis for measuring other constructs, such as creativity, communication and leadership, provided that it is possible to identify critical incidents that relate to the construct being assessed (Kyllonen and Lee 2005).

Assessing Dynamics

A fourth aspect of computer-based assessment is the possibility of not only assessing more than an answer or a product but also using information about the process involved to provide an assessment. This information is based on the analysis of times and sequences in data records in logs that track students' paths through a task, their choices of which material to access and decisions about when to start writing an answer (M. Ainley 2006; Hadwin et al. 2005). M. Ainley draws attention to two issues associated with the use of time trace data: the reliability and validity of single-item measures (which are necessarily the basis of trace records) and appropriate analytic methods for data that span a whole task and use the trend, continuities, discontinuities and contingencies in those data. Kyllonen (2009) identifies two other approaches to assessment that make use of time records available from computer-based assessments. One studies the times taken to complete tasks. The other uses the time spent in choosing between pairs of options to provide an assessment of attitudes or preferences, as in the *Implicit Association Test* (IAT).

Implementing Technology-Based Assessment

Technology-Based Assessments in Australia

Australian education systems, in successive iterations of the *National Goals for Schooling* (MCEETYA 1999, 2008), have placed considerable emphasis on the application of ICT in education. The national goals adopted in 1999 stated that when students leave school, they should 'be confident, creative and productive users of new technologies, particularly information and communication technologies, and understand the impact of those technologies on society' (MCEETYA 1999).

This was reiterated in the more recent *Declaration on Educational Goals for Young Australians,* which asserted that ‘in this digital age young people need to be highly skilled in the use of ICT’ (MCEECDYA 2008).

The implementation of ICT in education was guided by a plan entitled *Learning in an On-line World* (MCEETYA 2000, 2005) and supported by the establishment of a national company (*education.au*) to operate a resource network (*Education Network Australia* or *EdNA*) and a venture called the *Learning Federation* to develop digital learning objects for use in schools. More recently, the *Digital Education Revolution* (DER) has been included as a feature of the *National Education Reform Agenda* which is adding impetus to the use of ICT in education through support for improving ICT resources in schools, enhanced Internet connectivity and building programmes of teacher professional learning. Part of the context for these developments is the extent to which young people in Australia have access to and use ICT (and Web-based technology in particular) at home and at school. Australian teenagers continue to have access to, and use, ICT to a greater extent than their peers in most other countries and are among the highest users of ICT in the OECD (Anderson and Ainley 2010). It is also evident that Australian teachers (at least, teachers of mathematics and science in lower secondary school) are among the highest users of ICT in teaching (Ainley et al. 2009).

In 2005, Australia began a cycle of 3-yearly national surveys of the ICT literacy of students (MCEETYA 2007). Prior to the 2005 national assessment, the Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) defined ICT as the technologies used for accessing, gathering, manipulation and presentation or communication of information and adopted a definition of ICT Literacy as: *the ability of individuals to use ICT appropriately to access, manage, integrate and evaluate information, develop new understandings, and communicate with others in order to participate effectively in society* (MCEETYA 2007). This definition draws heavily on the Framework for ICT Literacy developed by the International ICT Literacy Panel and the OECD PISA ICT Literacy Feasibility Study (International ICT Literacy Panel 2002). ICT literacy is increasingly regarded as a broad set of generalizable and transferable knowledge, skills and understandings that are used to manage and communicate the cross-disciplinary commodity that is information. The integration of information and process is seen to transcend the application of ICT within any single learning discipline (Markauskaite 2007). Common to information literacy are the processes of identifying information needs, searching for and locating information and evaluating its quality, as well as transforming information and using it to communicate ideas (Catts and Lau 2008). According to Catts and Lau (2008), ‘people can be information literate in the absence of ICT, but the volume and variable quality of digital information, and its role in knowledge societies, has highlighted the need for all people to achieve information literacy skills’.

The Australian assessment framework envisaged ICT literacy as comprising six key processes: accessing information (identifying information requirements and knowing how to find and retrieve information); managing information (organizing and storing information for retrieval and reuse); evaluating (reflecting on the processes used to design and construct ICT solutions and judgments regarding the integrity, relevance and usefulness of information); developing new understandings (creating information and knowledge by synthesizing, adapting, applying, designing, inventing or authoring); communicating (exchanging information by sharing knowledge and creating information products to suit the audience, the context and the medium) and using ICT appropriately (critical, reflective and strategic ICT decisions and consideration of social, legal and ethical issues). Progress was envisaged in terms of levels of increasing complexity and sophistication in three strands of ICT use: (a) working with information, (b) creating and sharing information and (c) using ICT responsibly. In Working with Information, students progress from using keywords to retrieve information from a specified source, through identifying search question terms and suitable sources, to using a range of specialized sourcing tools and seeking confirmation of the credibility of information from external sources. In Creating and Sharing Information, students progress from using functions within software to edit, format, adapt and generate work for a specific purpose, through integrating and interpreting information from multiple sources with the selection and combination of software and tools, to using specialized tools to control, expand and author information, producing representations of complex phenomena. In Using ICT Responsibly, students progress from understanding and using basic terminology and uses of ICT in everyday life, through recognizing responsible use of ICT in particular contexts, to understanding the impact and influence of ICT over time and the social, economic and ethical issues associated with its use. These results can inform the refinement of a development progression of the type discussed in Chap. 3.

In the assessment, students completed all tasks on the computer by using a seamless combination of simulated and live software applications[1]. The tasks were grouped in thematically linked modules, each of which followed a linear narrative sequence. The narrative sequence in each module typically involved students collecting and appraising information before synthesizing and reframing it to suit a particular communicative purpose and given software genre. The overarching narratives across the modules covered a range of school-based and out-of-school-based themes. The assessment included items (such as simulated software operations) that were automatically scored and items that required constructed responses stored as text or as authentic software artefacts. The constructed response texts and artefacts were marked by human assessors.

[1] The assessment instrument integrated software from four different providers on a Microsoft Windows XT platform. The two key components of the software package were developed by SkillCheck Inc. (Boston, MA) and SoNet Software (Melbourne, Australia). The SkillCheck system provided the software responsible for delivering the assessment items and capturing student data. The SkillCheck system also provided the simulation, short constructed response and multiple-choice item platforms. The SoNet software enabled live software applications (such as Microsoft Word) to be run within the global assessment environment and for the resultant student products to be saved for later grading.

All students first completed a General Skills Test and then two randomly assigned (grade-appropriate) thematic modules. One reason for conducting the assessment with a number of modules was to ensure that the assessment instrument accessed what was common to the ICT Literacy construct across a sufficient breadth of contexts.

The modules followed a basic structure in which simulation, multiple-choice and short-constructed response items led up to a single large task using at least one live software application. Typically, the lead-up tasks required students to manage files, perform simple software functions (such as inserting pictures into files), search for information, collect and collate information, evaluate and analyse information and perform some simple reshaping of information (such as drawing a chart to represent numerical data). The large tasks that provided the global purpose of the modules were then completed using live software. When completing the large tasks, students typically needed to select, assimilate and synthesize the information they had been working with in the lead-up tasks and reframe it to fulfil a specified communicative purpose. Students spent between 40% and 50% of the time allocated for the module on the large task. The modules, with the associated tasks, were:

- Flag Design (Grade 6). Students use purpose-built previously unseen flag design graphics software to create a flag.
- Photo Album (Grades 6 and 10). Students use unseen photo album software to create a photo album to convince their cousin to come on holiday with them.
- DVD Day (Grades 6 and 10). Students navigate a closed Web environment to find information and complete a report template.
- Conservation Project (Grades 6 and 10). Students navigate a closed Web environment and use information provided in a spreadsheet to complete a report to the principal using Word.
- Video Games and Violence (Grade 10). Students use information provided as text and empirical data to create a PowerPoint presentation for their class.
- Help Desk (Grades 6 and 10). Students play the role of providing general advice on a community Help Desk and complete some formatting tasks in Word, PowerPoint and Excel.

The ICT literacy assessment was administered in a computer environment using sets of six networked laptop computers with all necessary software installed. A total of 3,746 grade 6 and 3,647 grade 10 students completed the survey in 263 elementary and 257 secondary schools across Australia. The assessment model defined a single variable, ICT literacy, which integrated three related strands. The calibration provided a high person separation index of 0.93 and a difference in the mean grade 6 ability compared to the mean grade 10 ability of the order of 1.7 logits, meaning that the assessment materials worked well in measuring individual students and in revealing differences associated with a developmental progression.

Describing the scale of achievement involved a detailed expert analysis of the ICT skills and knowledge required to achieve each score level on each item in the empirical scale. Each item, or partial credit item category, was then added to the empirical item scale to generate a detailed, descriptive ICT literacy scale. Descriptions were completed to describe the substantive ICT literacy content within each level.

At the bottom level (1), student performance was described as: Students perform basic tasks using computers and software. They implement the most commonly used file management and software commands when instructed. They recognize the most commonly used ICT terminology and functions.

At the middle level (3), students working at level 3 generate simple general search questions and select the best information source to meet a specific purpose. They retrieve information from given electronic sources to answer specific, concrete questions. They assemble information in a provided simple linear order to create information products. They use conventionally recognized software commands to edit and reformat information products. They recognize common examples in which ICT misuse may occur and suggest ways of avoiding them.

At the second top level (5), students working at level 5 evaluate the credibility of information from electronic sources and select the most relevant information to use for a specific communicative purpose. They create information products that show evidence of planning and technical competence. They use software features to reshape and present information graphically consistent with presentation conventions. They design information products that combine different elements and accurately represent their source data. They use available software features to enhance the appearance of their information products.

In addition to providing an assessment of ICT literacy, the national survey gathered information about a range of students' social characteristics and their access to ICT resources. There was a significant difference according to family socioeconomic status, with students whose parents were senior managers and professionals scoring rather higher than those whose parents were unskilled manual and office workers. Aboriginal and Torres Strait Islander students scored lower than other students. There was also a significant difference by geographic location. Allowing for all these differences in background, it was found that computer familiarity was an influence on ICT literacy. There was a net difference associated with frequency of computer use and with length of time for which computers had been used.

The assessment instrument used in 2008 was linked to that used in 2005 by the inclusion of three common modules (including the general skills test), but four new modules were added. The new modules included tasks associated with more interactive forms of communication and more extensively assessed issues involving responsible use. In addition, the application's functions were based on OpenOffice.

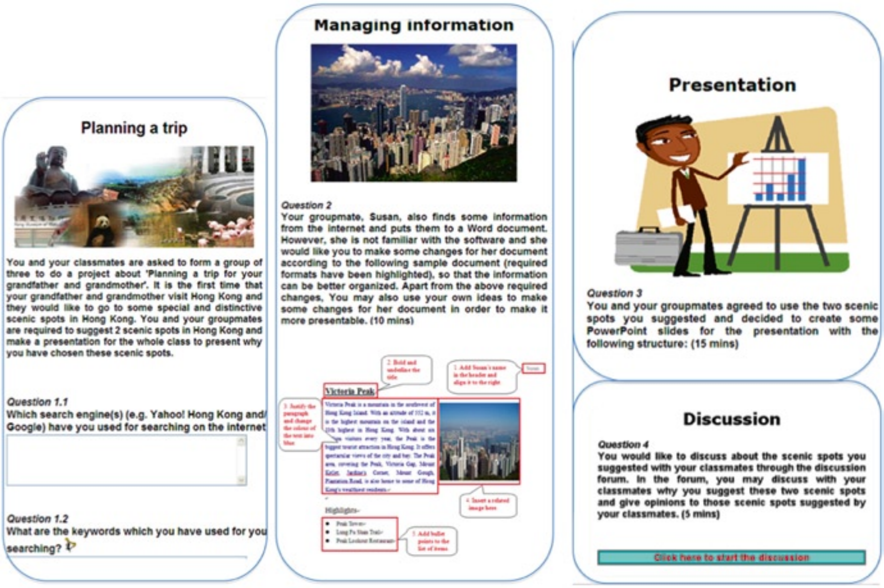
Technology-Based Assessments in Asia

In the major economies in Asia, there has been a strong move towards curriculum and pedagogical changes for preparing students for the knowledge economy since the turn of the millennium (Plomp et al. 2009). For example, 'Thinking Schools, Learning Nation' was the educational focus for Singapore's first IT in Education Masterplan (Singapore MOE 1997). The Hong Kong SAR government launched a comprehensive curriculum reform in 2000 (EMB 2001) focusing on developing students' lifelong learning capacity, which is also the focus of Japan's e-learning strategy (Sakayauchi et al. 2009). Pelgrum (2008) reports a shift in reported pedagogical practice from traditional towards twenty-first-century orientation in these countries between 1998 and 2006, which may reflect the impact of implementation of education policy in these countries.

The focus on innovation in curriculum and pedagogy in these Asian economies may have been accompanied by changes in the focus and format in assessment practice, including high-stakes examinations. For example, in Hong Kong, a teacher-assessed year-long independent enquiry is being introduced in the compulsory subject Liberal Studies, which forms 20% of the subject score in the school-leaving diploma at the end of grade 12 and is included in the application for university admission. This new form of assessment is designed to measure the generic skills that are considered important for the twenty-first century. On the other hand, technology-based means of assessment delivery have not been a focus of development in any of the Asian countries at the system level, although there may have been small-scale explorations by individual researchers. Technology-based assessment innovation is rare; one instance is the project on performance assessment of students' information literacy skills conducted in Hong Kong in 2007 as part of the evaluation of the second IT in education strategy in Hong Kong (Law et al. 2007, 2009). This project on Information Literacy Performance Assessment (ILPA for short, see http://il.cite.hku.hk/index.php) is described in some detail here as it attempts to use technology in the fourth and fifth domains of assessment described in an earlier section (whether someone is capable of achieving a higher level of performance with the appropriate use of general or domain-specific technology tools, and the ability to use technology to support collaboration and knowledge building).

Within the framework of the ILPA project, ICT literacy (IL) is not equated to technical competence. In other words, merely being technologically confident does not automatically lead to critical and skilful use of information. Technical know-how is inadequate by itself; individuals must possess the cognitive skills needed to identify and address various information needs and problems. ICT literacy includes both cognitive and technical proficiency. Cognitive proficiency refers to the desired foundational skills of everyday life at school, at home and at work. Seven information literacy dimensions were included in the assessment:

- Define—Using ICT tools to identify and appropriately represent information needs
- Access—Collecting and/or retrieving information in digital environments
- Manage—Using ICT tools to apply an existing organizational or classification scheme for information
- Integrate—Interpreting and representing information, such as by using ICT tools to synthesize, summarize, compare and contrast information from multiple sources
- Create—Adapting, applying, designing or inventing information in ICT environments
- Communicate—Communicating information properly in its context (audience and media) in ICT environments



Planning a trip

You and your classmates are asked to form a group of three to do a project about 'Planning a trip for your grandfather and grandmother'. It is the first time that your grandfather and grandmother visit Hong Kong and they would like to go to some special and distinctive scenic spots in Hong Kong. You and your groupmates are required to suggest 2 scenic spots in Hong Kong and make a presentation for the whole class to present why you have chosen these scenic spots.

Question 1.1
Which search engine(s) (e.g. Yahoo! Hong Kong and/ Google) have you used for searching on the internet

Question 1.2
What are the keywords which you have used for you searching?

Managing information

Question 2
Your groupmate, Susan, also finds some information from the internet and puts them to a Word document. However, she is not familiar with the software and she would like you to make some changes for her document according to the following sample document (required formats have been highlighted), so that the information can be better organized. Apart from the above required changes, You may also use your own ideas to make some changes for her document in order to make it more presentable. (10 mins)

Susan
1. Add Susan's name in the header and align it to the right
2. Bold and underline the title
Victoria Peak
3. Justify the paragraph and change the colour of the text into blue
Victoria Peak is a mountain in the southwest of Hong Kong Island. With an altitude of 552 m, it is the highest mountain on the island and the 31th highest in Hong Kong. With about six million visitors every year, the Peak is the biggest tourist attraction in Hong Kong. It offers spectacular views of the city and bay. The Peak area, covering the Peak, Victoria Gap, Mount Kellett, Jardine's Corner, Mount Gough, Plantation Road, is also home to some of Hong Kong's wealthiest residents.
4. Insert a related image here
Highlights
• Peak Tower
• Lung Fu Shan Trail
• Peak Lookout Restaurant
5. Add bullet points to the list of items

Presentation

Question 3
You and your groupmates agreed to use the two scenic spots you suggested and decided to create some PowerPoint slides for the presentation with the following structure: (15 mins)

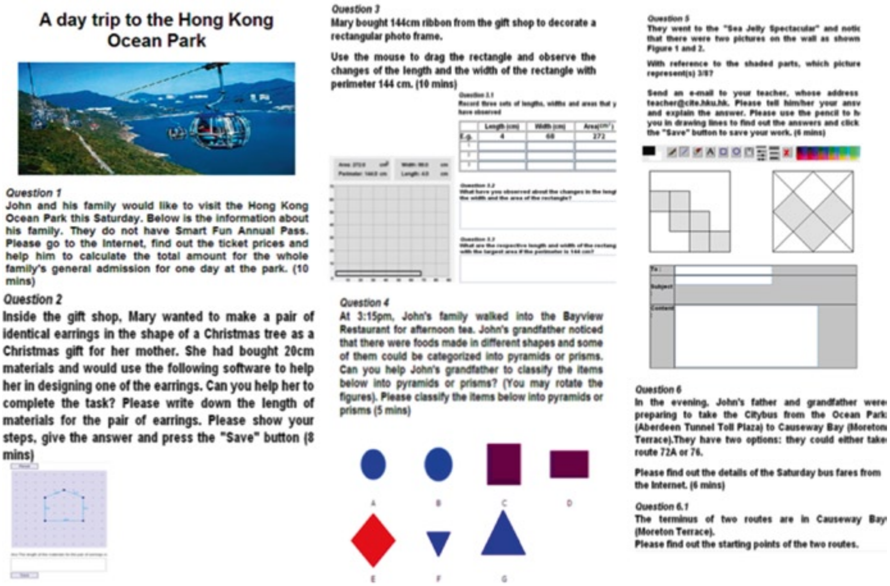
Discussion

Question 4
You would like to discuss about the scenic spots you suggested with your classmates through the discussion forum. In the forum, you may discuss with your classmates why you suggest these two scenic spots and give opinions to those scenic spots suggested by your classmates. (5 mins)

Fig. 4.1 Overview of performance assessment items for technical literacy (grades 5 and 8)

- Evaluate—Judging the degree to which information satisfies the needs of the task in ICT environments, including determining the authority, bias and timeliness of materials

While these dimensions are generic, a student's IL achievement is expected to be dependent on the subject matter domain context in which the assessment is conducted since the tools and problems may be very different. In this Hong Kong study, the target population participating in the assessment included primary 5 (P5, equivalent to grade 5) and secondary 2 (S2, equivalent to grade 8) students in the 2006/2007 academic year. Three performance assessments were designed and administered at each of these two grade levels. At P5, the assessments administered were a generic technical literacy assessment, IL in Chinese language and IL in mathematics. At S2, they were a generic technical literacy assessment, IL in Chinese language and IL in science. The generic technical literacy assessment tasks were designed to be the same at P5 and S2 levels as it was expected that personal and family background characteristics may have a stronger influence on a student's technical literacy than age. The assessment tasks for IL in Chinese language were designed to be different as the language literacy for these two levels of students was quite different. Overview of the performance assessments for technical literacy is presented in Fig. 4.1, that for information literacy in mathematics at grade 5 is presented in Fig. 4.2 and the corresponding assessment for information literacy in science at grade 8, in Fig. 4.3. It can be seen from these overviews that the tasks are



A day trip to the Hong Kong Ocean Park

Question 1
John and his family would like to visit the Hong Kong Ocean Park this Saturday. Below is the information about his family. They do not have Smart Fun Annual Pass. Please go to the Internet, find out the ticket prices and help him to calculate the total amount for the whole family's general admission for one day at the park. (10 mins)

Question 2
Inside the gift shop, Mary wanted to make a pair of identical earrings in the shape of a Christmas tree as a Christmas gift for her mother. She had bought 20cm materials and would use the following software to help her in designing one of the earrings. Can you help her to complete the task? Please write down the length of materials for the pair of earrings. Please show your steps, give the answer and press the "Save" button (8 mins)

Question 3
Mary bought 144cm ribbon from the gift shop to decorate a rectangular photo frame.

Use the mouse to drag the rectangle and observe the changes of the length and the width of the rectangle with perimeter 144 cm. (10 mins)

Question 3.1
Record three sets of lengths, widths and areas that y have observed

	Length (cm)	Width (cm)	Area(cm²)
E.g.	4	68	272
1			
2			
3			

Question 3.2
What have you observed about the changes in the lengt the width and the area of the rectangle?

Question 3.3
What are the respective length and width of the rectang with the largest area if the perimeter is 144 cm?

Question 4
At 3:15pm, John's family walked into the Bayview Restaurant for afternoon tea. John's grandfather noticed that there were foods made in different shapes and some of them could be categorized into pyramids or prisms. Can you help John's grandfather to classify the items below into pyramids or prisms? (You may rotate the figures). Please classify the items below into pyramids or prisms (5 mins)

Question 5
They went to the "Sea Jelly Spectacular" and notic that there were two pictures on the wall as shown Figure 1 and 2.

With reference to the shaded parts, which picture represent(s) 3/8?

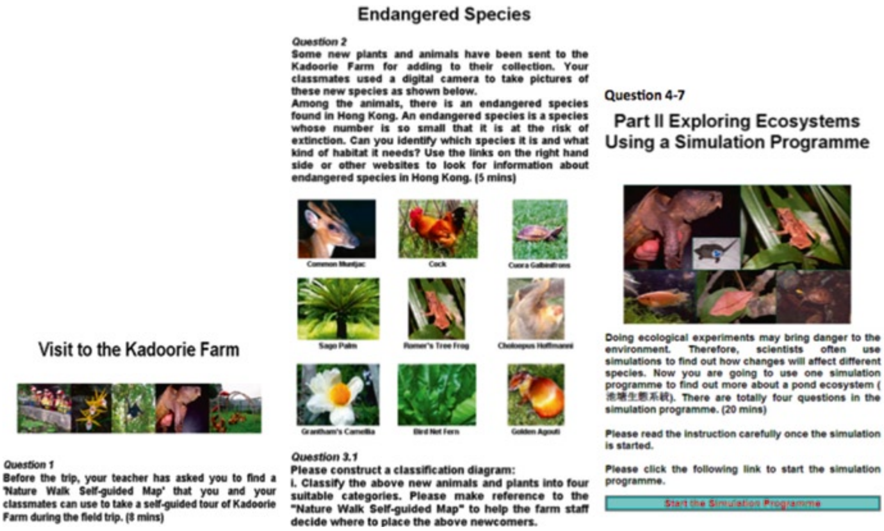
Send an e-mail to your teacher, whose address teacher@cite.hku.hk. Please tell him/her your ansv and explain the answer. Please use the pencil to h you in drawing lines to find out the answers and click the "Save" button to save your work. (6 mins)

Question 6
In the evening, John's father and grandfather were preparing to take the Citybus from the Ocean Park (Aberdeen Tunnel Toll Plaza) to Causeway Bay (Moreton Terrace).They have two options: they could either take route 72A or 76.

Please find out the details of the Saturday bus fares from the Internet. (6 mins)

Question 6.1
The terminus of two routes are in Causeway Bay (Moreton Terrace).
Please find out the starting points of the two routes.

Fig. 4.2 Overview of grade 5 performance assessment items for information literacy in mathematics



Visit to the Kadoorie Farm

Question 1
Before the trip, your teacher has asked you to find a 'Nature Walk Self-guided Map' that you and your classmates can use to take a self-guided tour of Kadoorie Farm during the field trip. (8 mins)

Endangered Species

Question 2
Some new plants and animals have been sent to the Kadoorie Farm for adding to their collection. Your classmates used a digital camera to take pictures of these new species as shown below.
Among the animals, there is an endangered species found in Hong Kong. An endangered species is a species whose number is so small that it is at the risk of extinction. Can you identify which species it is and what kind of habitat it needs? Use the links on the right hand side or other websites to look for information about endangered species in Hong Kong. (5 mins)

Question 3.1
Please construct a classification diagram:
i. Classify the above new animals and plants into four suitable categories. Please make reference to the "Nature Walk Self-guided Map" to help the farm staff decide where to place the above newcomers.

Question 4-7

Part II Exploring Ecosystems Using a Simulation Programme

Doing ecological experiments may bring danger to the environment. Therefore, scientists often use simulations to find out how changes will affect different species. Now you are going to use one simulation programme to find out more about a pond ecosystem (池塘生態系統). There are totally four questions in the simulation programme. (20 mins)

Please read the instruction carefully once the simulation is started.

Please click the following link to start the simulation programme.

Fig. 4.3 Overview of grade 8 performance assessment items for information literacy in science

designed to be authentic, i.e. related to everyday problems that students can understand and care about. Also, subject-specific tools are included; for instance, tools to support geometrical manipulation and tools for scientific simulation are included for the assessments in mathematics and science, respectively.

Since the use of technology is crucial to the assessment of information literacy, decisions on what kind of technology and how it is deployed in the performance assessment process are critical. It is important to ensure that students in all schools can have access to a uniform computing environment for the valid comparison of achievement in performance tasks involving the use of ICT. All primary and secondary schools in Hong Kong have at least one computer laboratory where all machines are connected to the Internet. However, the capability, age and condition of the computers in those laboratories differ enormously across different schools. The assumption of a computer platform that is generic enough to ensure that the educational applications designed can actually be installed in all schools is virtually impossible because of the complexity and diversity of ICT infrastructure in local schools. This problem is further aggravated by the lack of technical expertise in some schools such that there are often a lot of restrictions imposed on the functionalities available to students, such as disabling the right-click function, which makes some educational applications non-operable, and the absence of common plug-ins and applications, such as Active-X and Java runtime engines, so that many educational applications cannot be executed. In addition, many technical assistants are not able to identify problems to troubleshoot when difficulties occur.

The need for uniformity is particularly acute for the assessment of students' task performance using a variety of digital tools. Without a uniform technology platform in terms of the network connections and tools available, it is not possible to conduct fair assessment of students' performance, a task that is becoming increasingly important for providing authentic assessment of students' ability to perform tasks in the different subject areas that can make use of digital technology. Also, conducting the assessment in the students' own school setting was considered an important requirement as the study also wanted this experience to inform school-based performance assessment.

In order to solve this problem, the project team decided, after much exploration, on the use of a remote server system—the Microsoft Windows Terminal Server (WTS). This requires the computers in participating schools to be used only as thin clients, i.e. dumb terminals, during the assessment process, and it provides a unique and identical Windows' environment for every single user. Every computer in each participating school can log into the system and be used in the same way. In short, all the operations are independent for each client user, and functionalities are managed from the server operating system. Students and teachers can take part in learning sessions, surveys or assessments at any time and anywhere without worrying about the configurations of the computers on which they work. In addition to independent self-learning, collaborative learning with discussion can also be conducted within the WTS. While this set-up worked in many of the school sites, there were still a lot of technical challenges when the assessment was actually conducted, particularly issues related to firewall settings and bandwidth in schools.

All student actions during the assessment process were logged, and all their answers were stored on the server. Objective answers were automatically scored, while open-ended answers and digital artefacts produced by students were scored online, based on a carefully prepared and validated rubric that describes the performance observed at each level of achievement by experienced teachers in the relevant subject domains. Details of the findings are reported in Law et al. (2009).

Examples of Research and Development on Technology-Based Assessments in Europe

Using technology to make assessment more efficient is receiving growing attention in several European countries, and a research and development unit of the European Union is also facilitating these attempts by coordinating efforts and organizing workshops (Scheuermann and Björnsson 2009; Scheuermann and Pereira 2008).

At national level, Luxembourg has led the way by introducing a nationwide assessment system, moving immediately to online testing, while skipping the paper-based step. The current version of the system is able to assess an entire cohort simultaneously. It includes an advanced statistical analysis unit and the automatic generation of feedback to the teachers (Plichart et al. 2004, 2008). Created, developed and maintained in Luxembourg by the University of Luxembourg and the Public Research Center Henri Tudor, the core of the TAO (the acronym for Testing Assisté par Ordinateur, the French expression for Computer-Based Testing) platform has also been used in several international assessment programmes, including the Electronic Reading Assessment (ERA) in PISA 2009 (OECD 2008a) and the OECD Programme for International Assessment of Adult Competencies (PIAAC) (OECD 2008b). To fulfil the needs of the PIAAC household survey, computer-assisted personal interview (CAPI) functionalities have been fully integrated into the assessment capabilities. Several countries have also specialized similarly and further developed extension components that integrate with the TAO platform.

In Germany, a research unit of the *Deutsches Institut für Internationale Pädagogische Forschung* (DIPF, German Institute for International Educational Research, Frankfurt) has launched a major project that adapts and further develops the TAO platform. 'The main objective of the "Technology Based Assessment" (TBA) project at the DIPF is to establish a national standard for technology-assisted testing on the basis of innovative research and development according to international standards as well as reliable service.'[2] The technological aspects of the developmental work include item-builder software, the creation of innovative item formats (e.g. complex and interactive contents), feedback routines and computerized adaptive testing and item banks. Another innovative application of TBA is the measurement of complex problem-solving abilities; related experiments began in the late 1990s, and a large-scale assessment was conducted in the framework of the German extension of PISA 2003. The core of the assessment software is a finite

[2] See http://www.tba.dipf.de/index.php?option=com_content&task=view&id=25&Itemid=33 for the mission statement of the research unit.

automaton, which can be easily scaled in terms of item difficulty and can be realized in a number of contexts (cover stories, 'skins'). This approach provided an instrument that measures a cognitive construct distinct from both analytical problem-solving and general intelligence (Wirth and Klieme 2003; Wirth and Funke 2005). The most recent and more sophisticated tool uses the MicroDYN approach, where the testee faces a dynamically changing environment (Blech and Funke 2005; Greiff and Funke 2008). One of the major educational research initiatives, the Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes,[3] also includes several TBA-related studies (e.g. dynamic problem-solving, dynamic testing and rule-based item generation).

In Hungary, the first major technology-based testing took place in 2008. An inductive reasoning test was administered to a large sample of seventh grade students both in paper-and-pencil version and online (using the TAO platform) to examine the media effects. The first results indicate that although the global achievements are highly correlated, there are items with significantly different difficulties in the two media and there are persons who are significantly better on one or other of the media (Csapó et al. 2009). In 2009, a large-scale project was launched to develop an online diagnostic assessment system for the first six grades of primary school in reading, mathematics and science. The project includes developing assessment frameworks, devising a large number of items both on paper and on computer, building item banks, using technologies for migrating items from paper to computer and research on comparing the achievements on the tests using different media.

Examples of Technology in Assessment in the USA

In the USA, there are many instances in which technology is being used in large-scale summative testing. At the primary and secondary levels, the largest technology-based testing programmes are the Measures of Academic Progress (Northwest Evaluation Association), the Virginia Standards of Learning tests (Virginia Department of Education) and the Oregon Assessment of Knowledge and Skills (Oregon Department of Education). The Measures of Academic Progress (MAP) is a computer-adaptive test series offered in reading, mathematics, language usage and science at the primary and secondary levels. MAP is used by thousands of school districts. The test is linked to a diagnostic framework, DesCartes, which anchors the MAP score scale in skill descriptions that are popular with teachers because they appear to offer formative information. The Virginia Standards of Learning (SOL) tests are a series of assessments that cover reading, mathematics, sciences and other subjects at the primary and secondary levels. Over 1.5 million SOL tests are taken online annually. The Oregon Assessment of Knowledge and Skills (OAKS) is an adaptive test in reading, mathematics and science in primary and secondary grades.

[3] See http://kompetenzmodelle.dipf.de/en/projects.

The OAKS is approved for use under *No Child Left Behind*, the only adaptive test reaching that status. OAKS and those of the Virginia SOL tests used for *NCLB* purposes have high stakes for schools because sanctions can be levied for persistently poor test performance. Some of the tests may also have considerable stakes for students, including those measures that factor into end-of-course grading, promotion or graduation decisions. MAP, OAKS and SOL online assessments are believed to be based exclusively on multiple-choice tests.

Online tests offered by the major test publishers, for what the publishers describe as formative assessment purposes, include Acuity (CTB/McGraw-Hill) and the PASeries (Pearson). Perhaps more aligned with current concepts of formative assessment are the Cognitive Tutors (Carnegie Learning). The Cognitive Tutors, which focus on algebra and geometry, present problems to students, use their responses to dynamically judge understanding and then adjust the instruction accordingly.

At the post-secondary level, ACCUPLACER (College Board) and COMPASS (ACT) are summative tests used for placing entering freshmen in developmental reading, writing and mathematics courses. All sections of the tests are adaptive, except for the essay, which is automatically scored. The tests have relatively low stakes for students. The Graduate Record Examinations (GRE) General Test (ETS), the Graduate Management Admission Test (GMAT) (GMAC) and the Test of English as a Foreign Language (TOEFL) iBT (ETS) are all offered on computer. All three summative tests are high-stakes ones used in educational admissions. Sections of the GRE and GMAT are multiple-choice, adaptive tests. The writing sections of all three tests include essays, which are scored automatically and as well by one or more human graders. The TOEFL iBT also has a constructed-response speaking section, with digitized recordings of examinee responses scored by human judges. A formative assessment, TOEFL Practice Online (ETS), includes speaking questions that are scored automatically.

Applying Technology in International Assessment Programmes

The large-scale international assessment programmes currently in operation have their origins in the formation of the *International Association for the Evaluation of Educational Achievement* (IEA) in 1958. The formation of the IEA arose from a desire to focus comparative education on the study of variations in educational outcomes, such as knowledge, understanding, attitude and participation, as well as the inputs to education and the organization of schooling. Most of the current large-scale international assessment programmes are conducted by the IEA and the *Organization for Economic Co-operation and Development* (OECD).

The IEA has conducted the *Trends in International Mathematics and Science Study* (TIMSS) at grade 4 and grade 8 levels every 4 years since 1995 and has its fifth cycle scheduled for 2011 (Mullis et al. 2008; Martin et al. 2008). It has also conducted the *Progress in International Reading Literacy Study* (PIRLS) at grade 4 level every 5 years since 2001 and has its third cycle scheduled for 2011 (Mullis et al. 2007). In addition, the IEA has conducted periodic assessments in *Civic and*

Citizenship Education (ICCS) in 1999 (Torney-Purta et al. 2001) and 2009 (Schulz et al. 2008) and is planning an assessment of *Computer and Information Literacy* (ICILS) for 2013.

The OECD has conducted the *Programme for International Student Assessment* (PISA) among 15-year-old students every 3 years since 2000 and has its fifth cycle scheduled for 2012 (OECD 2007). It assesses reading, mathematical and scientific literacy in each cycle but with one of those three as the major domain in each cycle. In the 2003 cycle, it included an assessment of problem-solving. The OECD is also planning to conduct the *Programme for the International Assessment for Adult Competencies* (PIAAC) in 2011 in 27 countries. The target population is adults aged between 16 and 65 years, and each national sample will be a minimum of 5,000 people, who will be surveyed in their homes (OECD 2008b). It is designed to assess literacy, numeracy and 'problem solving skills in technology-rich environments,' as well as to survey how those skills are used at home, at work and in the community.

TIMSS and PIRLS have made use of ICT for Web-based school and teacher surveys but have not yet made extensive use of ICT for student assessment. An international option of Web-based reading was planned to be part of PIRLS 2011, and modules were developed and piloted. Whether the option proceeds to the main survey will depend upon the number of countries that opt to include the module. The *International Computer and Information Literacy Study* (ICILS) is examining the outcomes of student computer and information literacy (CIL) education across countries. It will investigate the variation in CIL outcomes between countries and between schools within countries so that those variations can be related to the way CIL education is provided. CIL is envisaged as the capacity to use computers to investigate, create and communicate in order to participate effectively at home, at school, in the workplace and in the community. It brings together computer competence and information literacy and envisages the strands of accessing and evaluating information, as well as producing and exchanging information. In addition to a computer-based student assessment, the study includes computer-based student, teacher and school surveys. It also incorporates a national context survey.

PISA has begun to use ICT in the assessment of the domains it assesses. In 2006, for PISA, scientific literacy was the major domain, and the assessment included an international option entitled a *Computer-Based Assessment of Science* (CBAS). CBAS was delivered by a test administrator taking a set of six laptop computers to each school, with the assessment system installed on a wireless or cabled network, with one of the networked PCs acting as an administrator's console (Haldane 2009). Student responses were saved during the test both on the student's computer and on the test administrator's computer. An online translation management system was developed to manage the translation and verification process for CBAS items. A typical CBAS item consisted of a stimulus area, containing text and a movie or flash animation, and a task area containing a simple or complex multiple-choice question, with radio buttons for selecting the answer(s). Some stimuli were interactive, with students able to set parameters by keying-in values or dragging scale pointers. There were a few drag-and-drop tasks, and some multiple-choice questions required students to select from a set of movies or animations. There were no constructed response items, all items were computer scored, and all student interactions with items were logged. CBAS field trials were conducted in 13 countries, but the option was included in the main study in only three of these.

PISA 2009 has reading literacy as a major domain and included *Electronic Reading Assessment* (ERA) as an international option. The ERA test uses a test administration system (TAO) developed through the University of Luxembourg (as described previously in this chapter). TAO can deliver tests over the Internet, across a network (as is the case with ERA) or on a stand-alone computer with student responses collected on a memory (Universal Serial Bus (USB)) stick. The ERA system includes an online translation management system and an online coding system for free-response items. An ERA item consists of a stimulus area that is a simulated multi-page Web environment and a task area. A typical ERA item involves students navigating around the Web environment to answer a multiple-choice or free-response question. Other types of tasks require students to interact in the stimulus area by clicking on a specific link, making a selection from a drop-down menu, posting a blog entry or typing an email. Answers to constructed-response items are collated to be marked by humans, while other tasks are scored by computer. The PISA 2009 *Reading Framework* articulates the constructs assessed in the ERA and the relationship of those constructs to the paper-based assessment. Subsequent cycles of PISA plan to make further use of computer-based assessment.

PIAAC builds on previous international surveys of adult literacy (such as IALS and ALL) but is extending the range of competencies assessed and investigating the way skills are used at work. Its assessment focus is on literacy, numeracy, reading components and 'problem-solving in technology-rich environments' (OECD 2008b), which refers to the cognitive skills required in the information age rather than computer skills and is similar to what is often called information literacy. This aspect of the assessment will focus on solving problems using multiple sources of information on a laptop computer. The problems are intended to involve accessing, evaluating, retrieving and processing information and incorporate technological and cognitive demands. The conceptions of literacy and numeracy in PIAAC emphasize competencies situated in a range of contexts as well as application, interpretation and communication. The term 'reading components' refers to basic skills, such as 'word recognition, decoding skills, vocabulary knowledge and fluency'. In addition to assessing these domains, PIAAC surveys adults in employment about the types and levels of a number of the general skills used in their workplaces, as well as background information, which includes data about how they use literacy, numeracy and technology skills in their daily lives, their education background, employment experience and demographic characteristics (OECD 2008b). The assessment, and the survey, is computer-based and administered to people in their homes by trained interviewers. The assessment is based on the TAO system.

In international assessment programmes, as in national and local programmes, two themes in the application of ICT are evident. One is the use of ICT to assess better the domains that have traditionally been the focus of assessment in schools: reading, mathematics and science. 'Assessing better' means using richer and more interactive assessment materials, using these materials to assess aspects of the domains that have been hard to assess and possibly extending the boundaries of those domains. This theme has been evident in the application of ICT thus far in PISA and PIRLS. A second theme is the use of ICT to assess more generic competencies. This is evident in the proposed ICILS and the PIAAC, which both propose to assess the use of computer technology to assess a broad set of generalizable and transferable knowledge, skills and understandings that are used to manage and communicate information. They are dealing with the intersection of technology and information literacy (Catts and Lau 2008).

Task Presentation, Response Capture and Scoring

Technological delivery can be designed to closely mimic the task presentation and response entry characteristics of conventional paper testing. Close imitation is important if the goal is to create a technology-delivered test capable of producing scores comparable to a paper version. If, however, no such restriction exists, technological delivery can be used to dramatically change task presentation, response capture and scoring.

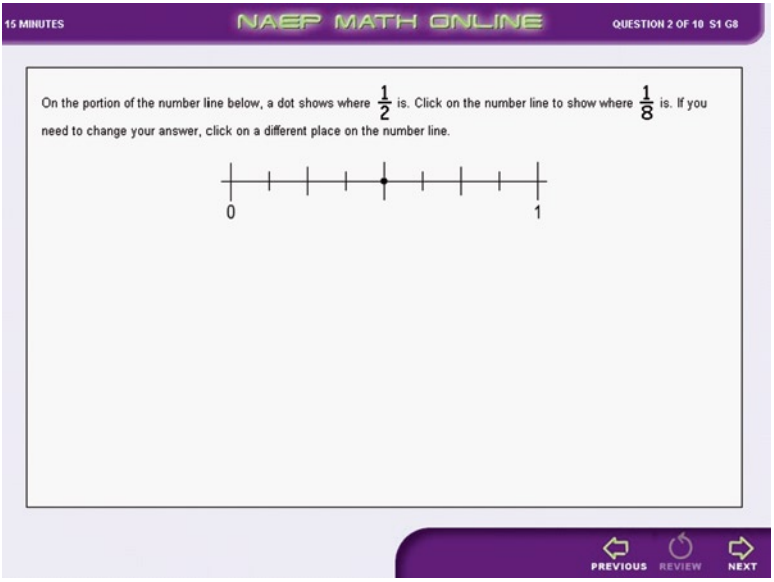
Task Presentation and Response Entry

Most technologically delivered tests administered today use traditional item types that call for the static presentation of a test question and the entry of a limited response, typically a mouse click in response to one of a small set of multiple-choice options. In some instances, test questions in current operational tests call for more elaborate responses, such as entering an essay.

In between a multiple-choice response and an elaborate response format, like an essay, there lies a large number of possibilities, and as has been a theme throughout this chapter, domain, purpose and context play a role in how those possibilities are implemented and where they might work most appropriately. Below, we give some examples for the three domain classes identified earlier: (1) domains in which practitioners interact with new technology primarily through the use of specialized tools, (2) domains in which technology may be used exclusively or not at all and (3) domains in which technology use is central to the definition.

Domains in Which Practitioners Primarily Use Specialized Tools

As noted earlier, in mathematics, students and practitioners tend to use technology tools for specialized purposes rather than pervasively in problem-solving. Because such specialized tools as spreadsheets and graphing calculators are not used generally, the measurement of students' mathematical skills on computer has



On the portion of the number line below, a dot shows where $\frac{1}{2}$ is. Click on the number line to show where $\frac{1}{8}$ is. If you need to change your answer, click on a different place on the number line.

Fig. 4.4 Inserting a point on a number line (Source: Bennett 2007)

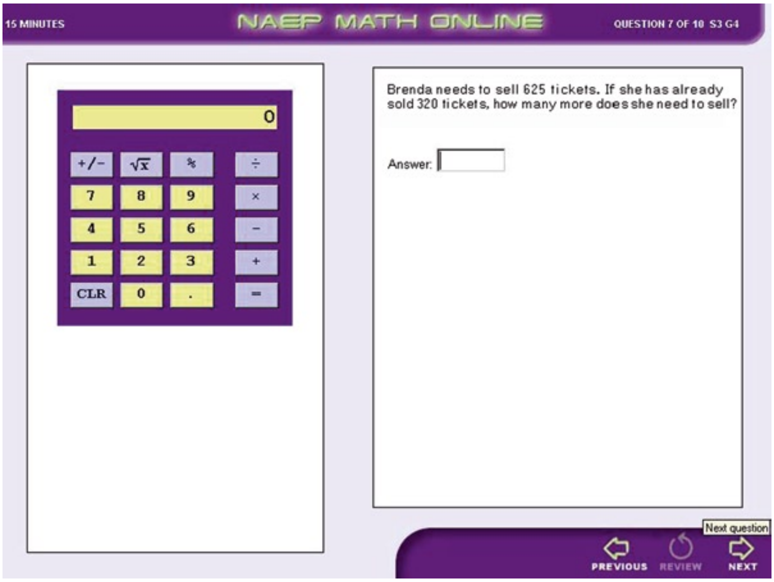
tended to track the manner of problem-solving as it is conventionally practised in classrooms and represented on paper tests, an approach which does not use the computer to maximum advantage. In this case, the computer serves primarily as a task presentation and response collection device, and the key goal is preventing the computer from becoming an impediment to problem-solving. That goal typically is achieved both through design and by affording students the opportunity to become familiar with testing on computer and the task formats. Developing that familiarity might best be done through formative assessment contexts that are low stakes for all concerned.

The examples presented in following figures illustrate the testing of mathematical competencies on computer that closely tracks the way those competencies are typically assessed on paper.

Figure 4.4 shows an example from a research study for the National Assessment of Educational Progress (NAEP) (Bennett 2007).

The task calls for the identification of a point on a number line that, on paper, would simply be marked by the student with a pencil. In this computer version, the student must use the mouse to click on the appropriate point on the line. Although this item format illustrates selecting from among choices, there is somewhat less of a forced-choice flavour than the typical multiple-option item because there are many more points from which to choose.

In Fig. 4.5, also from NAEP research, the examinee can use a calculator by clicking on the buttons, but must then enter a numeric answer in the response box. This process



Brenda needs to sell 625 tickets. If she has already sold 320 tickets, how many more does she need to sell?

Answer:

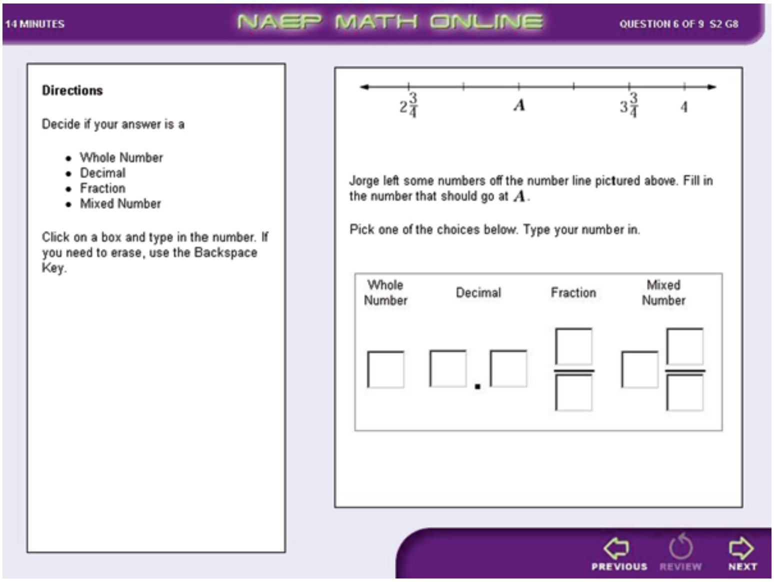
Fig. 4.5 A numeric entry task allowing use of an onscreen calculator (Source: Bennett 2007)

replicates what an examinee would do on a paper test using a physical calculator (compute the answer and then enter it onto the answer sheet). An alternative design for computer-based presentation would be to take the answer left in the calculator as the examinee's intended response to the problem.

An advantage in the use of an onscreen calculator is that the test developer controls when to make the calculator available to students (i.e. for all problems or for some subset). A second advantage is that the level of sophistication of the functions is also under the testing programme's control. Finally, all examinees have access to the same functions and must negotiate the same layout. To ensure that all students are familiar with that layout, some amount of practice prior to testing is necessary.

Figure 4.6 illustrates an instance from NAEP research in which the computer appeared to be an impediment to problem-solving. On paper, the item would simply require the student to enter a value into an empty box represented by the point on the number line designated by the letter 'A'. Implementing this item on computer raised the problem of how to insure that fractional responses were input in the mathematically preferred 'over/under' fashion while not cueing the student to the fact that the answer was a mixed number. This response type, however, turned what was a one-step problem on paper into a two-step problem on computer because the student had to choose the appropriate template before entering the response. The computer version of the problem proved to be considerably more difficult than the paper version (Sandene et al. 2005).

Figure 4.7 shows an example used in graduate admissions research (Bennett et al. 2000). Although requiring only the entry of numeric values, this response type



Directions

Decide if your answer is a

- Whole Number
- Decimal
- Fraction
- Mixed Number

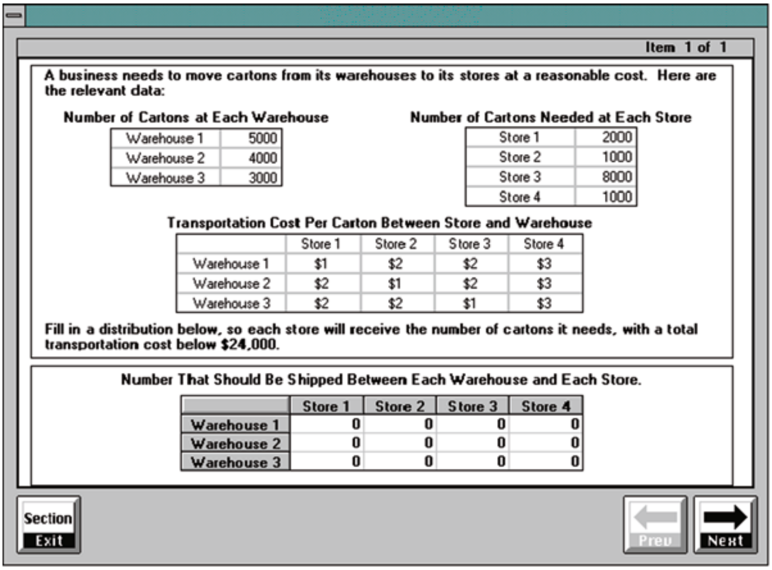
Click on a box and type in the number. If you need to erase, use the Backspace Key.

Jorge left some numbers off the number line pictured above. Fill in the number that should go at A.

Pick one of the choices below. Type your number in.

Whole Number	Decimal	Fraction	Mixed Number
□	□ . □	$\frac{□}{□}$	□ $\frac{□}{□}$

Fig. 4.6 A numeric entry task requiring use of a response template (Source: Bennett 2007)



Item 1 of 1

A business needs to move cartons from its warehouses to its stores at a reasonable cost. Here are the relevant data:

Number of Cartons at Each Warehouse

Warehouse 1	5000
Warehouse 2	4000
Warehouse 3	3000

Number of Cartons Needed at Each Store

Store 1	2000
Store 2	1000
Store 3	8000
Store 4	1000

Transportation Cost Per Carton Between Store and Warehouse

	Store 1	Store 2	Store 3	Store 4
Warehouse 1	$1	$2	$2	$3
Warehouse 2	$2	$1	$2	$3
Warehouse 3	$2	$2	$1	$3

Fill in a distribution below, so each store will receive the number of cartons it needs, with a total transportation cost below $24,000.

Number That Should Be Shipped Between Each Warehouse and Each Store.

	Store 1	Store 2	Store 3	Store 4
Warehouse 1	0	0	0	0
Warehouse 2	0	0	0	0
Warehouse 3	0	0	0	0

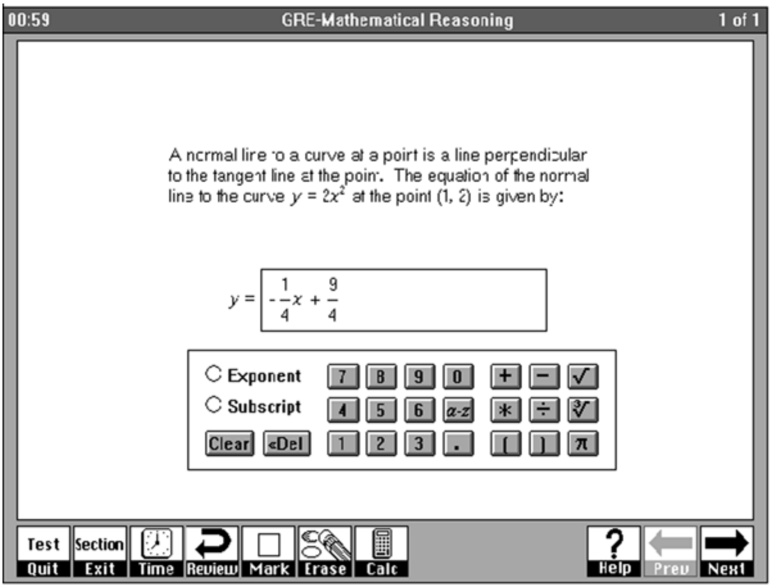
Fig. 4.7 Task with numeric entry and many correct answers to be scored automatically (Source: Bennett et al. (1998). Copyright (c) 1998 ETS. Used by permission)

is interesting for other reasons. The problem is cast in a business context. The stem gives three tables showing warehouses with inventory, stores with product needs and the costs associated with shipping between warehouses and stores, as well as an overall shipping budget. The task is to allocate the needed inventory to each store (using the bottom table) without exceeding the resources of the warehouses or the shipping budget.

The essence of this problem is *not* to find the best answer but only to find a reasonable one. Problems such as this one are typical of a large class of problems people encounter daily in real-world situations in which there are many right answers, the best answer may be too time consuming to find, and any of a large number of alternative solutions would be sufficient for many applied purposes.

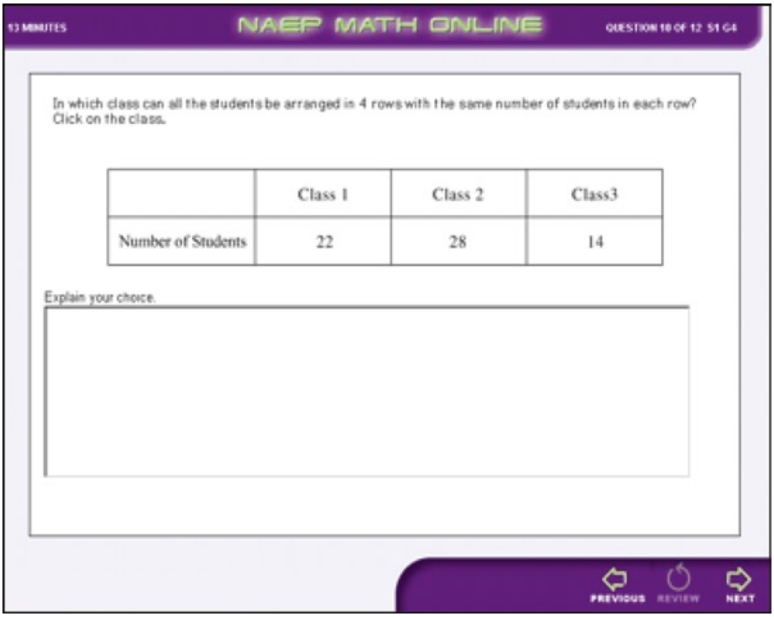
One attraction of presenting this type of item on computer is that even though there may be many correct answers, responses can be easily scored automatically. Scoring is done by testing each answer against the problem conditions. That is, does the student's answer fall within the resources of the warehouses, does it meet the stores' inventory needs, and does it satisfy the shipping budget? And, of course, many other problems with this same 'constraint-satisfaction' character can be created, all of which can be automatically scored.

Figure 4.8 shows another type used in graduate admissions research (Bennett et al. 2000). The response type allows questions that have symbolic expressions as answers, allowing, for example, situations presented as text or graphics to be modelled algebraically. To enter an expression, the examinee uses the mouse to click


A normal line to a curve at a point is a line perpendicular to the tangent line at the point. The equation of the normal line to the curve $y = 2x^2$ at the point (1, 2) is given by:

$y = -\frac{1}{4}x + \frac{9}{4}$

Fig. 4.8 Task requiring symbolic expression for answer (Source: Bennett et al. (1998). Copyright (c) 1998 ETS. Used by permission)



In which class can all the students be arranged in 4 rows with the same number of students in each row?
Click on the class.

	Class 1	Class 2	Class3
Number of Students	22	28	14

Explain your choice.

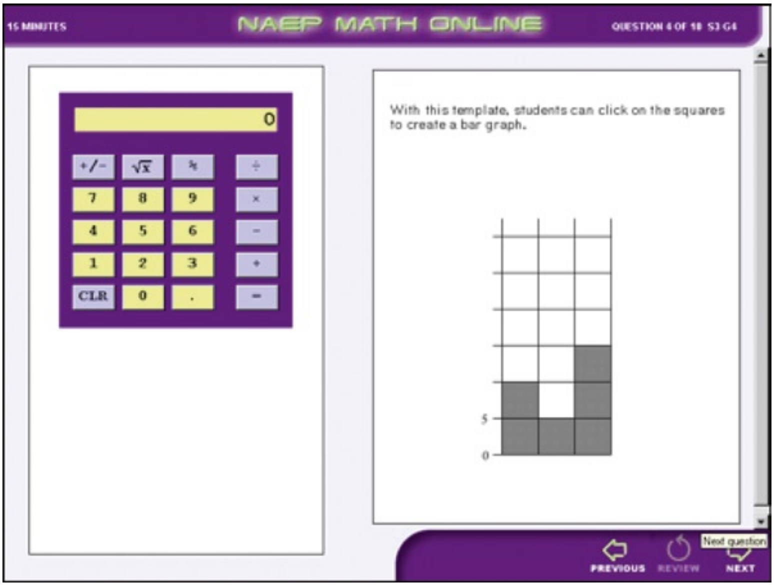
Fig. 4.9 Task requiring forced choice and text justification of choice (Source: Bennett 2007)

on the onscreen keypad. Response entry is not as simple as writing an expression on paper. In contrast to the NAEP format above, this response type avoids the need for multiple templates while still representing the response in over/under fashion. And, unlike paper, the responses can be automatically scored by testing whether the student's expression is algebraically equivalent to the test developer key.

In Fig. 4.9 is a question format from NAEP research in which the student must choose from among three options the class that has a number of students divisible by 4 and then enter text that justifies that answer. The written justification can be automatically scored but probably not as accurately as by human judges. Depending on the specific problem, the format might be used for gathering evidence related to whether a correct response indicates conceptual understanding or the level of critical thinking behind the answer choice.

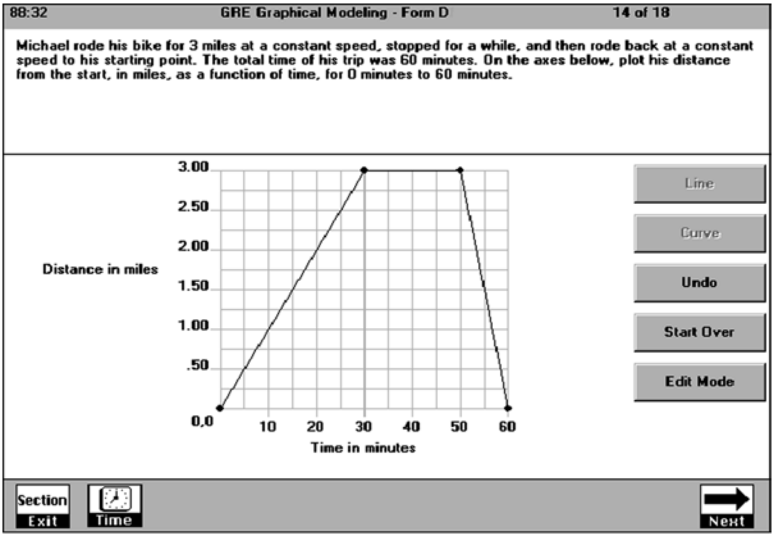
Figure 4.10 shows a NAEP-research format in which the student is given data and then must use the mouse to create a bar graph representing those data. Bars are created by clicking on cells in the grid to shade or unshade a box.

Figure 4.11 shows a more sophisticated graphing task used in graduate admissions research. Here, the examinee plots points on a grid and then connects them by pressing a line or curve button. With this response type, problems that have one correct answer or multiple correct answers can be presented, all of which can be scored automatically. In this particular instance, a correct answer is any trapezoidal shape like the one depicted that shows the start of the bicycle ride at 0 miles and 0 min; a stop almost any time at 3 miles and the conclusion at 0 miles and 60 min.



With this template, students can click on the squares to create a bar graph.

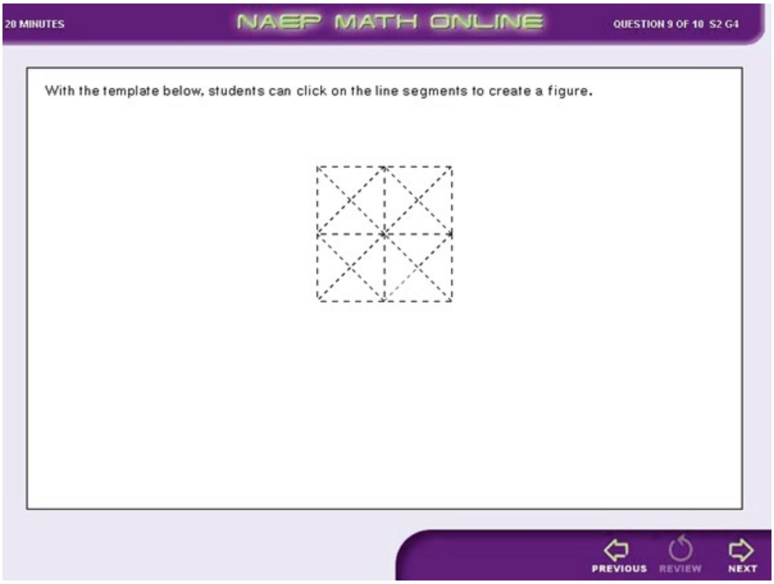
Fig. 4.10 Graph construction with mouse clicks to shade/unshade boxes (Source: Bennett 2007)



Michael rode his bike for 3 miles at a constant speed, stopped for a while, and then rode back at a constant speed to his starting point. The total time of his trip was 60 minutes. On the axes below, plot his distance from the start, in miles, as a function of time, for 0 minutes to 60 minutes.

Fig. 4.11 Plotting points on grid to create a line or curve (Source: Bennett et al. (1998). Copyright (c) 1998 ETS. Used by permission)

Finally, in the NAEP-research format shown in Fig. 4.12, the student is asked to create a geometric shape, say a right triangle, by clicking on the broken line segments, which become dark and continuous as soon as they are selected. The



With the template below, students can click on the line segments to create a figure.

Fig. 4.12 Item requiring construction of a geometric shape (Source: Bennett 2007)

advantage of this format over free-hand drawing, of course, is that the nature of the figure will be unambiguous and can be scored automatically.

In the response types above, the discussion has focused largely on the method of responding as the stimulus display itself differed in only limited ways from what might have been delivered in a paper test. And, indeed, the response types were generally modelled upon paper tests in an attempt to preserve comparability with problem-solving in that format.

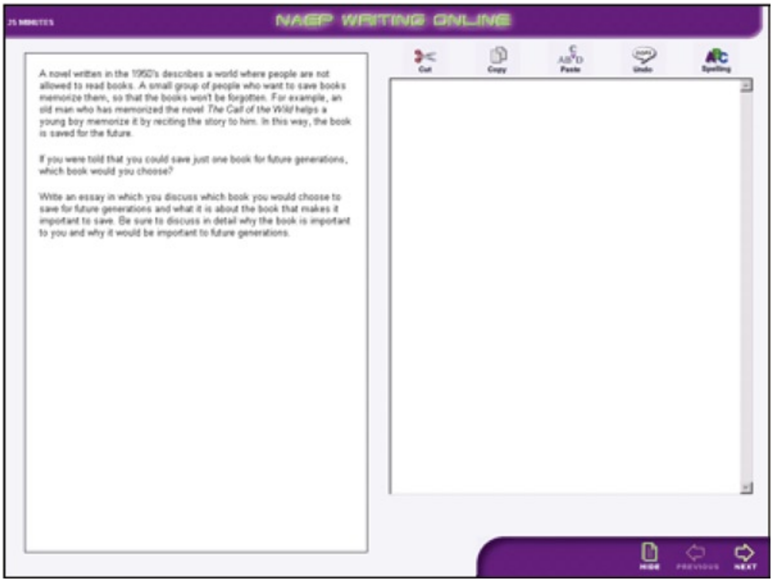
However, there are domains in which technology delivery can make the stimulus dynamic through the use of audio, video or animation, an effect that cannot be achieved in conventional tests unless special equipment is used (e.g. TV monitor with video playback). Listening comprehension is one such domain where, as in mathematics, interactive technology is not used pervasively in schools as part of the typical domain practice. For assessment purposes, dynamic presentation can be paired with traditional test questions, as when a student is presented with an audio clip from a lecture and then asked to respond onscreen to a multiple-choice question about the lecture. Tests like the TOEFL iBT (Test of English as a Foreign Language Internet-Based Test) pair such audio presentation with a still image, a choice that appears reasonable if the listening domain is intentionally conceptualized to exclude visual information. A more elaborate conception of the listening comprehension construct could be achieved if the use of visual cues is considered important by adding video of the speaker.

Science is a third instance in which interactive technology is not used pervasively in schools as part of the typical domain practice. Here, again, interactive tools are used for specialized purposes, such as spreadsheet modelling or running simulations of complex physical systems. Response formats used in testing might include responding to forced-choice and constructed-response questions after running simulated experiments or after observing dynamic phenomena presented in audio, video or animation.

There have been many notable projects that integrate the use of simulation and visualization tools to provide rich and authentic tasks for learning in science. Such learning environments facilitate a deeper understanding of complex relationships in many domains through interactive exploration (e.g. Mellar et al. 1994; Pea 2002; Feurzeig and Roberts 1999; Tinker and Xie 2008). Many of the technologies used in innovative science curricula also have the potential to be used or adapted for use in assessment in science education, opening up new possibilities for the kinds of student performances that can be examined for formative or summative purposes (Quellmalz and Haertel 2004). Some examples of the integration of such tools in assessment in science are given below to illustrate the range of situations and designs that can be found in the literature.

Among the earliest examples of technology-supported performance assessment in science that target non-traditional learning outcomes are the assessment tasks developed for the evaluation of the GLOBE environmental science education programme. One of the examples described by Means and Haertel (2002) was designed to measure inquiry skills associated with the analysis and interpretation of climate data. Here, students were presented with a set of climate-related criteria for selecting a site for the next Winter Olympics as well as multiple types of climate data on a number of possible candidate cities. The students had to analyse the sets of climate data using the given criteria, decide on the most suitable site on the basis of those results and then prepare a persuasive presentation incorporating displays of comparative climatic data to illustrate the reasons for their selection. The assessment was able to reveal the extent to which students were able to understand the criteria and to apply them consistently and systematically and whether they were able to present their argument in a clear and coherent manner. The assessment, therefore, served well its purpose of evaluating the GLOBE programme. However, Means and Haertel (2002) point out that as the assessment task was embedded within the learning system used in the programme, it could not be used to satisfy broader assessment needs. One of the ways they have explored for overcoming such limitations was the development and use of assessment templates to guide the design of classroom assessment tools.

The *SimScientists* assessment is a project that makes use of interactive simulation technology for the assessment of students' science learning outcomes, designed to support classroom formative assessment (Quellmalz and Pellegrino 2009; Quellmalz et al. 2009). The simulation-based assessments were designed according to an evidence-centred design model (Mislevy and Haertel 2006) such that the task designed will be based on models that elicit evidence of the targeted content and inquiry targets defined in the student model, and so the students' performance will be scored and reported on the basis of an appropriate evidence model for reporting on students' progress and achievement on the targets. In developing assessment tasks



A novel written in the 1950's describes a world where people are not allowed to read books. A small group of people who want to save books memorize them, so that the books won't be forgotten. For example, an old man who has memorized the novel *The Call of the Wild* helps a young boy memorize it by reciting the story to him. In this way, the book is saved for the future.

If you were told that you could save just one book for future generations, which book would you choose?

Write an essay in which you discuss which book you would choose to save for future generations and what it is about the book that makes it important to save. Be sure to discuss in detail why the book is important to you and why it would be important to future generations.

Fig. 4.13 A response type for essay writing (Source: Horkay and et al. 2005)

for specific content and inquiry targets, much attention is given to the identification of major misconceptions reported in the science education research literature that are related to the assessment targets as the assessment tasks are designed to reveal incorrect or naïve understanding. The assessment tasks are designed as formative resources by providing: (1) immediate feedback according to the students' performance, (2) real-time graduated coaching support to the student and (3) diagnostic information that can be used for further offline guidance and extension activities.

Domains in Which Technology Is Used Exclusively or Not at All

In the domain of writing, many individuals use the computer almost exclusively, while many others use it rarely or never. This situation has unique implications for design since the needs of both types of individuals must be accommodated in assessing writing.

Figure 4.13 shows an example format from NAEP research. On the left is writing prompt, and on the right is a response area that is like a simplified word processor. Six functions are available through tool buttons above the response area, including cutting, copying and pasting text; undoing the last action and checking spelling. Several of these functions are also accessible through standard keystroke combinations, like Control-C for copying text.

This format was intended to be familiar enough in its design and features to allow those proficient in writing on a computer to quickly and easily learn to use it, almost as they would in their typical writing activities. All the same, the design could work to the disadvantage of students who routinely use the more sophisticated features of commercial word processors.

The simple design of this response type was also intended to benefit those individuals who do not write on the computer at all. However, they would likely be disadvantaged by any design requiring keyboard input since computer familiarity, and particularly keyboarding skill, appears to affect online writing performance (Horkay et al. 2006). A more robust test design might also allow for handwritten input via a stylus. But even that input would require prior practice for those individuals not familiar with using a tablet computer. The essential point is that, for domains where some individuals practise primarily with technology tools and others do not, both forms of assessment, technology-delivered and traditional, may be necessary.

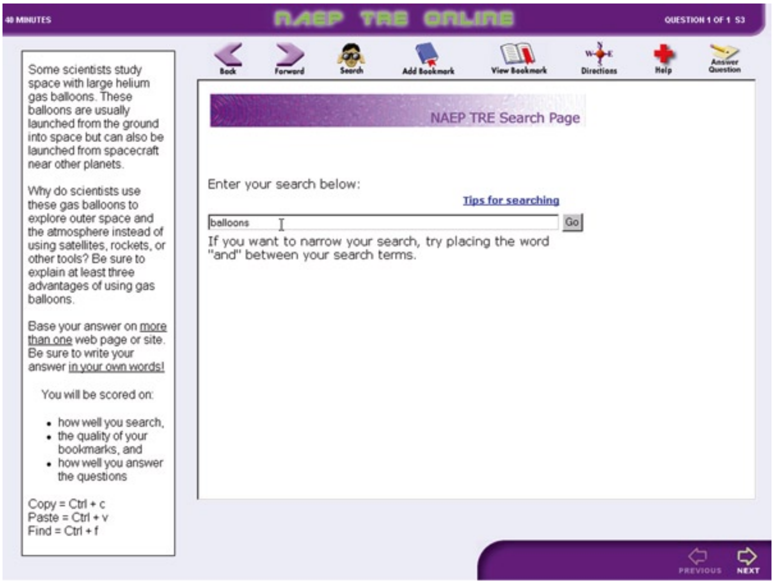
In assessment of writing, as in other domains where a technological tool is employed, a key issue is whether to create a simplified version of the tool for use in the assessment or to use the actual tool. Using the actual tool—in this instance, a particular commercial word processor—typically involves the substantial cost of licensing the technology (unless students use their own or institutional copies). That tool may also only run locally, making direct capture of response data by the testing agency more difficult. Third, if a particular word processor is chosen, this may advantage those students who use it routinely and disadvantage those who are used to a competitive product. Finally, it may not be easy, or even possible, to capture process data.

At the same time, there are issues associated with creating a generic tool, including decisions on what features to include in its design, the substantial cost of and time needed for development, and the fact that all students will need time to familiarize themselves with the resulting tool.

Domains in Which Technology Use Is Central to the Domain Definition

Technology-based assessment can probably realize its potential most fully and rapidly in domains where the use of interactive technology is central to the domain definition. In such domains, neither the practice nor the assessment can be done meaningfully without the technology. Although it can be used in either of the other two domain classes described above, simulation is a key tool in this third class of domains because it can be used to replicate the essential features of a particular technology or technology environment within which to assess domain proficiency.

An example can be found in the domain of electronic information search. Figure 4.14 shows a screen from a simulated Internet created for use in NAEP research (Bennett et al. 2007). On the left side of the screen is a problem statement, which asks the student to find out and explain why scientists sometimes use helium gas balloons for planetary atmospheric exploration. Below the problem statement is a summary of directions students have seen in more detail on previous screens. To the right is a search browser. Above the browser are buttons for revisiting pages,



Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets.

Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.

Base your answer on more than one web page or site. Be sure to write your answer in your own words!

You will be scored on:

- how well you search,
- the quality of your bookmarks, and
- how well you answer the questions

Copy = Ctrl + c
Paste = Ctrl + v
Find = Ctrl + f

NAEP TRE Search Page

Enter your search below:

Tips for searching

balloons

If you want to narrow your search, try placing the word "and" between your search terms.

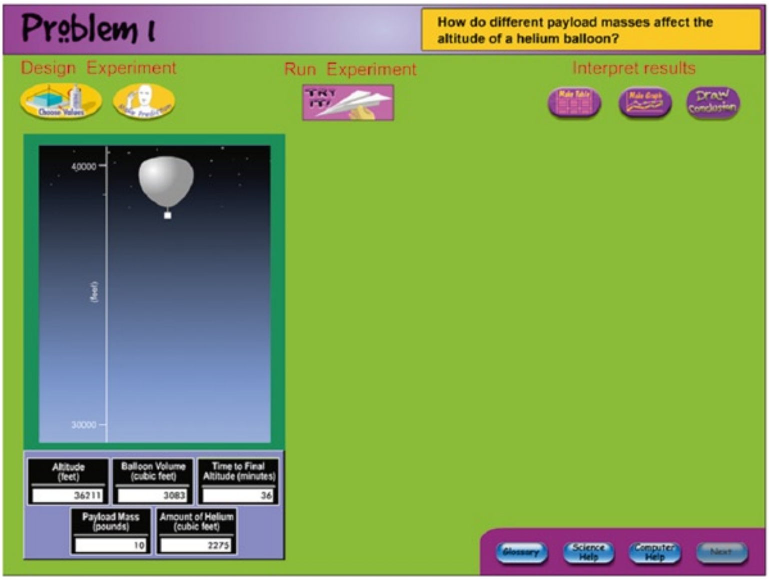
Fig. 4.14 A simulated Internet search problem (Source: Adapted from Bennett and et al. 2007)

bookmarking, going to the more extensive set of directions, getting hints and switching to a form to take notes or write an extended response to the question posed.

The database constructed to populate this simulated Internet consisted of some 5,000 pages taken from the real Internet, including pages devoted to both relevant and irrelevant material. A simulated Internet was used to ensure standardization because, depending upon school technology policy and the time of any given test administration, different portions of the real Internet could be available to students and it was necessary to prevent access to inappropriate sites from occurring under the auspices of NAEP. Each page in the database was rated for relevance to the question posed by one or more raters. To answer the set question, students had to visit multiple pages in the database and synthesize their findings. Student performance was scored both on the quality of the answer written in response to the question and on the basis of search behaviour. Among other things, the use of advanced search techniques like quotes, or the NOT operator, the use of bookmarks, the relevance of the pages visited or bookmarked and the number of searches required to produce a set of relevant hits were all factored into the scoring.

Of particular note is that the exercise will unfold differently, depending upon the actions the examinee takes—upon the number and content of search queries entered and the particular pages visited. In that sense, the problem will not be the same for all students.

A second example comes from the use of simulation for conducting experiments. In addition to the electronic information-search exercise shown earlier, Bennett et al. (2007) created an environment in which eighth grade students were asked to



How do different payload masses affect the altitude of a helium balloon?

Altitude (feet): 36211
Balloon Volume (cubic feet): 3083
Time to Final Altitude (minutes): 36
Payload Mass (pounds): 10
Amount of Helium (cubic feet): 2275

Fig. 4.15 Environment for problem-solving by conducting simulated experiments (Source: Adapted from Bennett and et al. 2007)

discover the relationships among various physical quantities by running simulated experiments. The experiments involved manipulating the payload mass carried by, and the amount of helium put into, a scientific gas balloon so as to determine the relationship of these variables with the altitude to which the balloon can rise in the atmosphere. The interface that the students worked with is shown in Fig. 4.15.

Depending on the specific problem presented (see upper right corner), the environment allows the student to select values for the independent variable of choice (payload mass and/or amount of helium), make predictions about what will happen to the balloon, launch the balloon, make a table or a graph and write an extended response to the problem. Students may go through the problem-solving process in any order and may conduct as many experiments as they wish. The behaviour of the balloon is depicted dynamically in the flight window and on the instrument panel below, which gives its altitude, volume, time to final altitude, payload mass carried and amount of helium put into it. Student performance was scored on the basis of the accuracy and completeness of the written response to the problem and upon aspects of the process used in solution. Those aspects included whether the number of experiments and range of the independent variable covered were sufficient to discover the relationship of interest, whether tables or graphs that incorporated all variables pertinent to the problem were constructed and whether the experiments were controlled so that the effects of different independent variables could be isolated.

Scoring

For multiple-choice questions, the scoring technology is well established. For constructed-response question types, including some of those illustrated above, the technology for machine scoring is only just emerging. Drasgow, Luecht and Bennett (2006) describe three classes of automated scoring of constructed response.

The first class is defined by a simple match between the scoring key and the examinee response. The response type given in Fig. 4.4 (requiring the selection of a point on a number line) would fall into this class, as would a reading passage that asks a student to click on the point at which a given sentence should be inserted, problems that call for ordering numerical values by dragging and dropping them into slots, extending a bar on a chart to represent a particular amount or entering a numeric response. In general, responses like these can be scored objectively. For some of these instances, tolerances for making fine distinctions in scoring need to be set. As an example, if a question directs the examinee to click on the point on the number line represented by 2.5 and the interface allows clicks to be made anywhere on the line, some degree of latitude in what constitutes a correct response will need to be permitted. Alternatively, the response type can be configured to accept only clicks at certain intervals.

A second problem class concerns what Drasgow et al. term static ones too complex to be graded by simple match. These problems are static in the sense that the task remains the same regardless of the actions taken by the student. Examples from this class include mathematical questions calling for the entry of expressions (Fig. 4.8), points plotted on a coordinate plane (Fig. 4.11) or numeric entries to questions having multiple correct answers (Fig. 4.7). Other examples are problems requiring a short written response, a concept map, an essay or a speech sample. Considerable work has been done on this category of automated scoring, especially for essays (Shermis and Burstein 2003), and such scoring is used operationally for summative assessment purposes that have high stakes for individuals by several large testing programmes, including the Graduate Record Examinations (GRE) General Test, the Graduate Management Admission Test (GMAT) and the TOEFL iBT. The automated scoring of low-entropy (highly predictable) speech is also beginning to see use in summative testing applications as well as that for less predictable, high-entropy speech in low-stakes, formative assessment contexts (Xi et al. 2008).

The third class of problems covers those instances in which the problem changes as a function of the actions the examinee takes in the course of solution. The electronic-search response type shown in Fig. 4.14 falls into this class. These problems usually require significant time for examinees to complete, and due to their highly interactive nature, they produce extensive amounts of data; every keystroke, mouse click and resulting event can be captured. Those facts suggest the need, also the opportunity, to use more than a correct end result as evidence for overall proficiency and further to pull out dimensions in addition to an overall proficiency. Achieving these goals, however, has proven to be exceedingly difficult since inevitably only some of the reams of data produced may be relevant. Deciding what to capture and what to score should be based upon a careful analysis of the domain conceptualization and the claims one wishes to make about examinees, the behaviours that would provide evidence for those claims and the tasks that will provide that evidence (Mislevy et al. 2004; Mislevy et al. 2006). Approaches to the scoring of problems in this class have been demonstrated for strategy use in scientific problem-solving (Stevens et al. 1996; Stevens and Casillas 2006), problem-solving with technology (Bennett et al. 2003), patient management for medical licensure (Clyman et al. 1995) and computer network troubleshooting (Williamson et al. 2006a, b).

For all three classes of constructed response, and for forced-choice questions too, computer delivery offers an additional piece of information not captured by a paper test—timing. That information may involve only the simple latency of the response for multiple-choice questions and constructed response questions in the first class (simple match) described above, where the simple latency is the time between the item's first presentation and the examinee's response entry. The timing data will be more complex for the second and third problem classes. An essay response, for example, permits latency data to be computed within and between words, sentences and paragraphs. Some of those latencies may have implications for measuring keyboard skills (e.g. within word), whereas others may be more suggestive of ideational fluency (e.g. between sentences).

The value of timing data will depend upon assessment domain, purpose and context. Among other things, timing information might be most appropriate for domains in which fluency and automaticity are critical (e.g. reading, decoding, basic number facts), for formative assessment purposes (e.g. where some types of delay may suggest the need for skill improvement) and when the test has low stakes for students (e.g. to determine which students are taking the test seriously).

Validity Issues Raised by the Use of Technology for Assessment

Below, we discuss several general validity issues, including some of the implications of the use of technology for assessment in the three domain classes identified earlier: (1) domains in which practitioners interact with new technology primarily through the use of specialized tools, (2) domains in which technology may be used exclusively or not at all and (3) domains in which technology use is central.

Chief among the threats to validity are (1) the extent to which an assessment fails to fully measure the construct of interest and (2) where other constructs tangential to the one of interest inadvertently influence test performance (Messick 1989). With respect to the first threat, no single response type can be expected to fully represent a complex construct, certainly not one as complex (and as yet undefined) as'twenty-first century skills'. Rather, each response type, and its method of scoring, should be evaluated theoretically and empirically with respect to the particular portion of the construct it represents. Ultimately, it is the complete measure itself, as an assembly of different response types, which needs to be subjected to evaluation of the extent to which it adequately represents the construct for some particular measurement purpose and context.

A particularly pertinent issue concerning construct representation and technology arises as a result of the advent of automated scoring (although it also occurs in human scoring). At a high level, automated scoring can be decomposed into three separable processes: feature extraction, feature evaluation and feature accumulation (Drasgow et al. 2006). Feature extraction involves isolating scorable components, feature evaluation entails judging those components, and feature accumulation consists of combining the judgments into a score or other characterization. In automated essay scoring, for example, a scorable component may be the discourse unit (e.g. introduction, body, conclusion), judged as present or absent, and then the number of these present, combined with similar judgments from other scorable components (e.g. average word complexity, average word length). The choice of the aspects of writing to score, how to judge these aspects and how to combine the judgments all bring into play concerns for construct representation. Automated scoring programmes, for example, tend to use features that are easily computable and to combine them in ways that best predict the scores awarded by human judges under operational conditions. Even when it predicts operational human scores reasonably well, such an approach may not provide the most effective representation of the writing construct (Bennett 2006; Bennett and Bejar 1998), omitting features that cannot be easily extracted from an essay by machine and, for the features that are extracted, giving undue weight to those that human experts would not necessarily value very highly (Ben-Simon and Bennett 2007).

The second threat, construct-irrelevant variance, also cannot be precisely identified in the absence of a clear definition of the construct of interest. Without knowing the exact target of measurement, it can be difficult to identify factors that might be irrelevant. Here, too, an evaluation can be conducted at the level of the response type as long as one can make some presumptions about what the test, overall, was *not* supposed to measure.

Construct under-representation and construct-irrelevant variance can be factored into a third consideration that is key to the measurement of domain classes 1 and 2, the comparability of scores between the conventional and technology-based forms of a test. Although different definitions exist, a common conceptualization is that scores may be considered comparable across two delivery modes when those modes produce highly similar rank orders of individuals and highly similar score distributions (APA 1986, p. 18). If the rank-ordering criterion is met but the distributions are not the same, it may be possible to make scores interchangeable through equating. Differences in rank order, however, are usually not salvageable through statistical adjustment. A finding of score comparability between two testing modes implies that the modes represent the construct equally well and that neither mode is differentially affected by construct-irrelevant variance. That said, such a finding indicates neither that the modes represent the construct sufficiently for a given purpose nor that they are uncontaminated by construct-irrelevant variance; it implies only that scores from the modes are equivalent—in whatever it is that they measure. Last, a finding that scores are not comparable suggests that the modes differ either in their degree of construct representation, in construct-irrelevant variance or both.

Comparability of scores across testing modes is important when a test is offered in two modes concurrently and users wish scores from the modes to be interchangeable.

Comparability may also be important when there is a transition from conventional to technological delivery and users wish to compare performance over time. There have been many studies of the comparability of paper and computer-based tests of cognitive skills for adults, leading to the general finding that scores are interchangeable for power tests but not for speeded measures (Mead and Drasgow 1993). In primary and secondary school populations, the situation is less certain (Drasgow et al. 2006). Several meta-analyses have concluded that achievement tests produce comparable scores (Kingston 2009; Wang et al. 2007, 2008). This conclusion, however, is best viewed as preliminary, because the summarized effects have come largely from: analyses of distribution differences with little consideration of rank-order differences; multiple-choice measures; unrepresentative samples; non-random assignment to modes; unpublished studies and a few investigators without accounting for violations of independence. In studies using nationally representative samples of middle-school students with random assignment to modes, analyses more sensitive to rank order and constructed-response items, the conclusion that scores are generally interchangeable across modes has not been supported (e.g. Bennett et al. 2008; Horkay et al. 2006).

It should be evident that, for domain class 3, score comparability across modes can play no role, because technology is central to the domain practice and, putatively, such practice cannot be measured effectively without using technology. For this domain class, only one testing mode should be offered. However, a set of claims about what the assessment is intended to measure and evidence about the extent to which those claims are supported is still essential, as it would be for any domain class. The claims and evidence needed to support validity take the form of an argument that includes theory, logic and empirical data (Kane 2006; Messick 1989).

For domain class 1, where individuals interact with technology primarily through the use of specialized tools, assessment programmes often choose to measure the entire domain on the computer even though some (or even most) of the domain components are not typically practised in a technology environment. This decision may be motivated by a desire for faster score turn-around or for other pragmatic reasons. For those domain components that are not typically practised on computer, construct-irrelevant variance may be introduced into problem-solving if the computer presentation used for assessment diverges too far from the typical domain (or classroom instructional) practice.

Figure 4.6 illustrates such an instance from NAEP mathematics research in which the computer appeared to be an impediment to problem-solving. In this problem, the student was asked to enter a value that represented a point on a number line. The computer version proved to be considerably more difficult than the paper version presumably because the former added a requirement not present in the paper mode (the need to select a response template before entering an answer) (Sandene et al. 2005). It is worth noting that this alleged source of irrelevant variance might have been trained away by sufficient practice with this response format in advance of the test. It is also worth noting that, under some circumstances, working with such a format might not be considered irrelevant at all (e.g. if such a template-selection procedure was typically used in mathematical problem-solving in the target population of students).

Figure 4.5 offers a second example. In this response type, created for use in graduate and professional admissions testing, the student enters complex expressions using a soft keypad (Bennett et al. 2000). Gallagher et al. (2002) administered problems using this response type to college seniors and first-year graduate students in mathematics-related fields. The focus of the study was to identify whether construct-irrelevant variance was associated with the response-entry process. Examinees were given parallel paper and computer mathematical tests, along with a test of expression editing and entry skill. The study found no mean score differences between the modes, similar rank orderings across modes and non-significant correlations of each mode with the edit-entry test (implying that among the range of editing-skill levels observed, editing skill made no difference in mathematical test score). However, 77% of examinees indicated that they would prefer to take the test on paper were it to count, with only 7% preferring the computer version. Further, a substantial portion mentioned having difficulty on the computer test with the response-entry procedure. The investigators then retrospectively sampled paper responses and tried to enter them on computer, finding that some paper responses proved too long to fit into the on-screen answer box, suggesting that some students might have tried to enter such expressions on the computer version but had to reformulate them to fit the required frame. If so, these students did their reformulations quickly enough to avoid a negative impact on their scores (which would have been detected by the statistical analysis). Even so, having to rethink and re-enter lengthy expressions was likely to have caused unnecessary stress and time pressure. For individuals less skilled with computer than these mathematically adept college seniors and first-year graduate students, the potential for irrelevant variance would seem considerably greater.

In the design of tests for domain classes 1 and 2, there might be instances where comparability is *not* expected because the different domain competencies are not intended to be measured across modes. For instance, in domain class 1, the conventional test may have been built to measure those domain components typically practised on paper while the technology test was built to tap primarily those domain components brought to bear when using specialized technology tools. In domain class 2, paper and computer versions of a test may be offered but, because those who practice the domain on paper may be unable to do so on computer (and vice versa), neither measurement of the same competencies nor comparable scores should be expected. This situation would appear to be the case in many countries among primary and secondary school students for summative writing assessments. Some students may be able to compose a timed response equally well in either mode but, as appeared to be the case for US eighth graders in NAEP research, many perform better in one or the other mode (Horkay et al. 2006). If student groups self-select to testing mode, differences in performance between the groups may become uninterpretable. Such differences could be the result of skill level (i.e. those who typically use one mode may be generally more skilled than those who typically use the other) or mode (e.g. one mode may offer features that aid performance in ways that the other mode does not) or else due to the interaction between the two (e.g. more skilled practitioners may benefit more from one mode than the other, while less skilled practitioners are affected equally by both modes).

An additional comparability issue relevant to computer-based tests *regardless* of domain class is the comparability of scores across hardware and software configurations, including between laptops and desktops, monitors of various sizes and resolutions and screen-refresh latencies (as may occur due to differences in Internet bandwidth). There has been very little recent published research on this issue but the studies that have been conducted suggest that such differences can affect score comparability (Bridgeman et al. 2003; Horkay et al. 2006). Bridgeman et al., for example, found reading comprehension scores to be higher for students taking a summative test on a larger, higher-resolution display than for students using a smaller, lower resolution screen. Horkay et al. found low-stakes summative test performance to be, in some cases, lower for students taking an essay test on a NAEP laptop than on their school computer, which was usually a desktop. Differences, for example, in keyboard and screen quality between desktops and laptops have greatly diminished over the past decade. However, the introduction of netbooks, with widely varying keyboards and displays, makes score comparability as a function of machine characteristics a continuing concern across domain classes.

Construct under-representation, construct-irrelevant variance and score comparability all relate to the meaning or scores or other characterizations (e.g. diagnostic statements) coming from an assessment. Some assessment purposes and contexts bring into play claims that require substantiation beyond that related to the meaning of these scores or characterizations. Such claims are implicit, or more appropriately explicit, in the theory of action that underlies use of the assessment (Kane 2006). A timely example is summative assessment such as that used under the US *No Child Left Behind* Act. Such summative assessment is intended not only to measure student (and group) standing, but explicitly to facilitate school improvement through various legally mandated, remedial actions. A second example is formative assessment in general. The claims underlying the use of such assessments are that they will promote greater achievement than would otherwise occur. In both the case of *NCLB* summative assessment and of formative assessment, evidence needs to be provided, first, to support the quality (i.e. validity, reliability and fairness) of the characterizations of students (or institutions) coming from the measurement instrument (or process). Such evidence is needed regardless of whether those characterizations are scores or qualitative descriptions (e.g. a qualitative description in the summative case would be, 'the student is proficient in reading; in the formative case, 'the student misunderstands borrowing in two-digit subtraction and needs targeted instruction on that concept'). Second, evidence needs to be provided to support the claims about the impact on individuals or institutions that the assessments are intended to have. Impact claims are the province of programme evaluation and relate to whether use of the assessment has had its intended effects on student learning or on other classroom or institutional practices. It is important to realize that evidence of impact is required *in addition to*, not as substitute for, evidence of score meaning, even for formative assessment purposes. Both types of evidence are required to support the validity and efficacy arguments that underlie assessments intended to effect change on individuals or institutions (Bennett 2009, pp. 14–17; Kane 2006, pp. 53–56).

One implication of this separation of score meaning and efficacy is that assessments delivered in multiple modes may differ in score meaning, in impact or in both. One could, for example, envision a formative assessment programme offered on both paper and computer whose characterizations of student understanding and of how to adapt instruction were equivalent—i.e. equally valid, reliable and fair—but that were differentially effective because the results of one were delivered faster than the results of the other.

Special Applications and Testing Situations Enabled by New Technologies

As has already been discussed in the previous sections, technology offers opportunities for assessment in domains and contexts where assessment would otherwise not be possible or would be difficult. Beyond extending the possibilities of routinely applied mainstream assessments, technology makes testing possible in several specific cases and situations. Two rapidly growing areas are discussed here; developments in both areas being driven by the needs of educational practices. Both areas of application still face several challenges, and exploiting the full potential of technology in these areas requires further research and developmental work.

Assessing Students with Special Educational Needs

For those students whose development is different from the typical, for whatever reason, there are strong tendencies in modern societies to teach them together with their peers. This is referred to as mainstreaming, inclusive education or integration—there are other terms. Furthermore, those who face challenges are provided with extra care and facilities to overcome their difficulties, following the principles of equal educational opportunities. Students who need this type of special care will be referred to here as students with Special Educational Needs (SEN). The definition of SEN students changes widely from country to country, so the proportion of SEN students within a population may vary over a broad range. Taking all kinds of special needs into account, in some countries this proportion may be up to 30%. This number indicates that using technology to assess SEN students is not a marginal issue and that using technology may vitally improve many students' chance for success in education and later for leading a complete life.

The availability of specially trained teachers and experts often limits the fulfilment of these educational ideals, but technology can often fill the gaps. In several cases, using technology instead of relying on the services of human helpers is not merely a replacement with limitations, but an enhancement of the personal capabilities of SEN students that makes independent learning possible.

In some cases, there may be a continuum between slow (but steady) development, temporal difficulties and specific developmental disorders. In other cases, development is severely hindered by specific factors; early identification and treatment of these may help to solve the problems. In the most severe cases, personal handicaps cannot be corrected, and technology is used to improve functionality.

As the inclusion of students with special educational needs in regular classrooms is an accepted basic practice, there is a growing demand for assessing together those students who are taught together (see Chap. 12 of Koretz 2008). Technology may be applied in this process in a number of different ways.

- Scalable fonts, using larger fonts.
- Speech synthesizers for reading texts.
- Blind students may enter responses to specific keywords.
- Development of a large number of specific technology-based diagnostic tests is in progress. TBA may reduce the need for specially trained experts and improve the precision of measurement, especially in the psychomotor area.
- Customized interfaces devised for physically handicapped students. From simple instruments to sophisticated eye tracking, these can make testing accessible for students with a broad range of physical handicaps (Lőrincz 2008).
- Adapting tests to the individual needs of students. The concept of adaptive testing may be generalized to identify some types of learning difficulties and to offer items matched to students' specific needs.
- Assessments built into specific technology-supported learning programmes. A reading improvement and speech therapy programme recognizes the intonation, the tempo and the loudness of speech or reading aloud and compares these to pre-recorded standards and provides visual feedback to students (http://www.inf.u-szeged.hu/beszedmester).

Today, these technologies are already available, and many of them are routinely used in e-learning (Ball et al. 2006; Reich and Petter 2009). However, transferring and implementing these technologies into the area of TBA requires further developmental work. Including SEN students in mainstream TBA assessment is, on the one hand, desirable, but measuring their achievements on the same scale raises several methodological and theoretical issues.

Connecting Individuals: Assessing Collaborative Skills and Group Achievement

Sfard (1998) distinguishes two main metaphors in learning: learning as acquisition and learning as participation. CSCL and collaborative learning, in general, belong more to the participation metaphor, which focuses on learning as becoming a participant, and interactions through discourse and activity as the key processes. Depending on the theory of learning underpinning the focus on collaboration, the learning outcomes to be assessed may be different (Dillenbourg et al. 1996). Assessing learning as an individual outcome is consistent with a socio-constructivist or socio-cultural view of learning, as social interaction provides conditions that are conducive to conflict resolution in learning (socio-constructivist) or scaffold learning through bridging the zone of proximal development (socio-cultural). On the other hand, a shared cognition approach to collaborative learning (Suchman 1987; Lave 1988) considers the learning context and environment as an integral part of the cognitive activity and a collaborating group can be seen as forming a single cognizing unit (Dillenbourg et al. 1996), and assessing learning beyond the individual poses an even bigger challenge.

Webb (1995) provides an in-depth discussion, based on a comprehensive review of studies on collaboration and learning, of the theoretical and practical challenges of assessing collaboration in large-scale assessment programmes. In particular, she highlights the importance of defining clearly the purpose of the assessment and giving serious consideration to the goal of group work and the group processes that are supposed to contribute to those goals to make sure that these work towards, rather than against, the purpose of the assessment. Three purposes of assessment were delineated in which collaboration plays an important part: the level of an individual's performance after learning through collaboration, group productivity and an individual's ability to interact and function effectively as a member of a team. Different assessment purposes entail different group tasks. Group processes leading to good performance are often different depending on the task and could even be competitive. For example, if the goal of the collaboration is group productivity, taking the time to explain to each other, so as to enhance individual learning through collaboration, may lower group productivity for a given period of time. The purpose of the assessment should also be made clear, as this will influence individual behaviour in the group. If the purpose is to measure individual student learning, Webb suggests that the test instructions should focus on individual accountability and individual performance in the group work and to include in the instruction what constitutes desirable group processes and why. On the other hand, a focus on group productivity may act against equality of participation and may even lead to a socio-dynamic in which low-status members' contributions are ignored. Webb's paper also reviewed studies on group composition (in terms of gender, personality, ability, etc.) and group productivity. The review clearly indicates that group composition is one of the important issues in large-scale assessments of collaboration.

Owing to the complexities in assessing cognitive outcomes in collaboration, global measures of participation such as frequency of response or the absence of disruptive behaviour are often used as indicators of collaboration, which falls far short of being able to reveal the much more nuanced learning outcomes such as the ability to explore a problem, generate a plan or design a product. Means et al. (2000) describe a Palm-top Collaboration Assessment project in which they developed an assessment tool that teachers can use for 'mobile real-time assessments' of collaboration skills as they move among groups of collaborating students. Teachers can use the tool to rate each group's performance on nine dimensions of collaboration (p.9):

- Analysing the Task
- Developing Social Norms
- Assigning and Adapting Roles
- Explaining/Forming Arguments

- Sharing Resources
- Asking Questions
- Transforming Participation
- Developing Shared Ideas and Understandings
- Presenting Findings

Teachers' ratings would be made on a three-point scale for each dimension and would be stored on the computer for subsequent review and processing.

Unfortunately, research that develops assessment tools and instruments independent of specific collaboration contexts such as the above is rare, even though studies of collaboration and CSCL are becoming an important area in educational research. On the other hand, much of the literature on assessing collaboration, whether computers are being used or not, is linked to research on collaborative learning contexts. These may be embedded as an integral part of the pedagogical design such as in peer- and self-assessment (e.g. Boud et al. 1999; McConnell 2002; Macdonald 2003), and the primary aim is to promote learning through collaboration. The focus of some studies involving assessment of collaboration is on the evaluation of specific pedagogical design principles. Lehtinen et al. (1999) summarizes the questions addressed in these kinds of studies as belonging to three different paradigms. 'Is collaborative learning more efficient than learning alone?' is typical of questions under the effects paradigm. Research within the conditions paradigm studies how learning outcomes are influenced by various conditions of collaboration such as group composition, task design, collaboration context and the communication/collaboration environment. There are also studies that examine group collaboration development in terms of stages of inquiry (e.g. Gunawardena et al. 1997), demonstration of critical thinking skills (e.g. Henri 1992) and stages in the development of a socio-metacognitive dynamic for knowledge building within groups engaging in collaborative inquiry (e.g. Law 2005).

In summary, in assessing collaboration, both the unit of assessment (individual or group) and the nature of the assessment goal (cognitive, metacognitive, social or task productivity) can be very different. This poses serious methodological challenges to what and how this is to be assessed. Technological considerations and design are subservient to these more holistic aspects in assessment.

Designing Technology-Based Assessment

Formalizing Descriptors for Technology-Based Assessment

Assessment in general and computer-based assessment in particular is characterized by a large number of variables that influence decisions on aspects of organization, methodology and technology. In turn these decisions strongly influence the level of risk and its management, change management, costs and timelines. Decisions on the global design of an evaluation programme can be considered as a bijection between the assessment characteristic space and the assessment design space ($D = C \otimes D$, $D = \{O,M,T\}$). In order to scope and address assessment challenges and better support decision-making, beyond the inherent characteristics of the framework and instrument themselves, one needs to define a series of dimensions describing the assessment space. It is not the purpose of this chapter to discuss thoroughly each of these dimensions and their relationship with technologies, methods, instruments and organizational processes. It is important, however, to describe briefly the most important features of assessment descriptors. A more detailed and integrated analysis should be undertaken to establish best practice recommendations. In addition to the above-mentioned descriptors, one can also cite those following.

Scale

The scale of an assessment should not be confused with its objective. Indeed, when considering assessment objectives, one considers the level of *granularity* of the relevant and meaningful information that is collected and analysed during the evaluation. Depending on the assessment object, the lowest level of granularity, the elementary piece of information, may either be individual scores or average scores over populations or sub-populations, considered as systems or sub-systems. The scale of the assessment depicts the *number* of information units collected, somewhat related to the size of the sample. Exams at school level and certification tests are typically small-scale assessments, while PISA or NAEP are typically large-scale operations.

Theoretical Grounds

This assessment descriptor corresponds to the theoretical framework used to set up the measurement scale. *Classical* assessment uses a (possibly weighted) ratio of correct answers to total number of questions while Item Response Theory (IRT) uses statistical parameterization of items. As a sub-descriptor, a scoring method must be considered from theoretical as well as procedural or algorithmic points of view.

Scoring Mode

Scoring of the items and of the entire test, in addition to reference models and procedures, can be *automatic*, *semi-automatic* or *manual*. Depending on this scoring mode, organizational processes and technological support, as well as risks to security and measurement quality, may change dramatically.

Reference

In some situations, the data collected does not reflect objective evidence of achievement on the scale or metrics. Subjective evaluations are based on test takers' assertions about their own level of achievement, or potentially, in the case of hetero-evaluation, about others' levels of achievement. These situations are referred to as declarative assessment, while scores inferred from facts and observations collected by an agent other than the test taker are referred to as evidence-based assessments.

Framework Type

Assessments are designed for different contexts and for different purposes on the basis of a reference description of the competency, skill or ability that one intends to measure. These various frameworks have different origins, among which the most important are *educational programmes and training specifications* (content-based or goal-oriented); *cognitive constructs* and *skill cards and job descriptions*. The type of framework may have strong implications for organizational processes, methodology and technical aspects of the instruments.

Technology Purpose

The function of technology in assessment operations is another very important factor that has an impact on the organizational, methodological and technological aspects of the assessment. While many variations can be observed, two typical situations can be identified: *computer-aided assessment* and *computer-based assessment*. In the former, the technology is essentially used at the level of organizational and operational support processes. The assessment instrument remains paper-and-pencil and IT is only used as a support tool for the survey. In the latter situation, the computer itself is used to deliver the instrument.

Context Variables

Depending on the scale of the survey, a series of scaling variables related to the context are also of great importance. Typical variables of this type are *multi-lingualism*; *multi-cultural aspects*; consideration of *disabilities*; *geographical aspects* (remoteness); *geopolitical, political and legal aspects*; *data collection mode* (e.g. centralized, network-based, in-house).

Stakeholders

The identification of the stakeholders and their characteristics is important for organizational, methodological and technological applications. Typical stakeholders are the *test taker*, the *test administrator* and the *test backer*.

Intentionality/Directionality

Depending on the roles and relationships between stakeholders, the assessment will require different intentions and risks to be managed. Typical situations can be described by asking two fundamental questions: (a) which stakeholder assigns the assessment to which other stakeholder? (b) which stakeholder evaluates which other stakeholder (in other words, which stakeholder provides the evidence or data collected during their assessment)? As an illustration this raises the notion of *self-assessment* where the test taker assigns a test to himself (be it declarative or evidence-based) and manipulates the instrument; or *hetero-assessment* (most generally declarative) where the respondent provides information to evaluate somebody else. In most classical situations, the test taker is different from the stakeholder who assigns the test.

Technology for Item Development and Test Management

One of the main success factors in developing a modern technology-based assessment platform is certainly not the level of technology alone; it relies on the adoption of an iterative and participatory design mode for the platform design and development process. Indeed, as is often observed in the field of scientific computing, the classical customer-supplier relationship that takes a purely Software Engineering service point of view is highly ineffective in such dramatically complex circumstances, in which computer science considerations are sometimes not separable from psychometric considerations. On the contrary, a successful technology-based assessment (TBA) expertise must be built on deep immersion in both disciplines.

In addition to the trans-disciplinary approach, two other factors will also increase the chance to fulfil the needs for the assessment of the twenty-first-century skills. First, the platform should be designed and implemented independently from any single specific context of use. This requires a more abstract level of design that leads to high-level and generic requirements that might appear remote from concrete user concepts or the pragmatics of organization. Consequently, a strong commitment and understanding on this issue by assessment experts together with a thorough understanding by technologists of the TBA domain, as well as good communication are essential. As already stressed in e-learning contexts, a strong collaboration between disciplines is essential (Corbiere 2008).

Secondly, TBA processes and requirements are highly multi-form and carry a tremendous diversity of needs and practices, not only in the education domain (Martin et al. 2009) but also more generally when ranging across assessment classification descriptors—from researchers in psychometrics, educational measurement or experimental psychology to large-scale assessment and monitoring professionals—or from the education context to human resource management. As a consequence, any willingness to build a comprehensive and detailed *a priori* description of the needs might appear totally elusive. Despite this, both assessment and technology experts should acknowledge the need to iteratively elicit the context-specific requirements that will be further abstracted in the analysis phase while the software is developed in a parallel process, in such a way that unexpected new features can be added with the least impact on the code. This process is likely to be the most efficient way to tackle the challenge.

Principles for Developing Technological Platforms

Enabling the Assessment of Reliability of Data and Versatility of Instruments

Instead of strongly depending on providers' business models, the open-source paradigm in this area bears two fundamental advantages. The full availability of the source code gives the possibility of assessing the implementation and reliability of the measurement instruments (a crucial aspect of scientific computing in general and psychometrics in particular). In addition it facilitates fine-tuning the software to very specific needs and contexts, keeping full control over the implementation process and costs while benefiting from the contributions of a possibly large community of users and developers (Latour & Farcot 2008). Built-in extension mechanisms enable developers from within the community to create new extensions and adaptations without modifying the core layers of the application and to share their contributions.

Enabling Efficient Management of Assessment Resources

An integrated technology-based assessment should enable the efficient management of assessment resources (items, tests, subjects and groups of subjects, results, surveys, deliveries and so on) and provide support to the organizational processes (depending on the context, translation and verification, for instance); the platform should also enable the delivery of the cognitive instruments and background questionnaires to the test takers and possibly other stakeholders, together with collecting, post-processing and exporting results and behavioural data. In order to support complex collaborative processes such as those needed in large-scale international surveys, a modern CBA platform should offer annotation with semantically rich meta-data as well as collaborative capabilities.

Complementary to the delivery of the cognitive Instruments, modern CBA platforms should also provide a full set of functionalities to collect background information, mostly about the test taker, but also possibly about any kind of resources involved in the process. As an example, in the PIAAC survey, a Background Questionnaire (consisting of questions, variables and logical flow of questions with branching rules) has been fully integrated into the global survey workflow, along with the cognitive instrument booklet.

In the ideal case, interview items, assessment items and entire tests or booklets are interchangeable. As a consequence, very complex assessment instruments can be designed to fully integrate cognitive assessment and background data collection in a single flow, on any specific platform.

Accommodating a Diversity of Assessment Situations

In order to accommodate the large diversity of assessment situations, modern computer-assessment platforms should offer a large set of deployment modes, from a fully Web-based installation on a large server-farm, with load balancing that enables the delivery of a large number of simultaneous tests, to distribution via CDs or memory sticks running on school desktops. As an illustration, the latter solution has been used in the PISA ERA 2009. In the PIAAC international survey, the deployment has been made using a Virtual Machine installed on individual laptops brought by interviewers into the participating households. In classroom contexts, wireless Local Area Network (LAN) using a simple laptop as server and tablet PC's as the client machines for the test takers can also be used.

Item Building Tools

Balancing Usability and Flexibility

Item authoring is one of the crucial tasks in the delivery of technology-based assessments. Up to the present, depending on the requirements of the frameworks, various strategies have been pursued, ranging from hard-coded development by software programmers to easy-to-use simple template-based authoring. Even if it seems intuitively to be the most natural solution, the purely programmer-provided process should in general be avoided. Such an outsourcing strategy (disconnecting the content specialists from the software developers) usually requires very precise specifications that item designers and framework experts are mostly not familiar with. In addition, it lengthens the timeline and reduces the number of iterations, preventing trial-and-error procedures. Moreover, this process does not scale well when the number of versions of every single item increases, as is the case when one has to deal with many languages and country-specific adaptations. Of course, there will always be a trade-off between usability and simplicity (that introduce strong constraints and low freedom in the item functionalities) and flexibility in describing rich interactive behaviours (that introduces a higher level of complexity when using the tool). In most situations, it is advisable to provide different interfaces dedicated to users with different levels of IT competency. To face the challenge of allowing great flexibility while keeping the system useable with a minimum of learning, template-driven authoring tools built on a generic expressive system are probably one of the most promising technologies. Indeed, this enables the use of a single system to hide inherent complexity when building simple items while giving more powerful users the possibility to further edit advanced features.

Separating Item Design and Implementation

Item-authoring processes can be further subdivided into the tasks of item design (setting up the item content, task definition, response domain and possibly scenarios) and item implementation (translating the item design for the computer platform, so that the item becomes an executable piece of software). Depending on the complexity of the framework, different tools can be used to perform each of the tasks. In some circumstances, building the items iteratively enables one to keep managing the items' complexity by first creating a document describing all the details of the item scenario, based on the framework definition, and then transforming it into an initial implementation template or draft. An IT specialist or a trained power user can then further expand the implementation draft to produce the executable form of the item. This process more effectively addresses stakeholders' requirements by remaining as close as possible to usual user practice. Indeed, modern Web- and XML-based technologies, such as CSS (Lie and Bos 2008), Javascript, HTML (Raggett et al. 1999), XSLT (Kay 2007), Xpath (Berglund et al. 2007) and Xtiger (Kia et al. 2008), among others, allow the easy building of template-driven authoring tools (Flores et al. 2006), letting the user having a similar experience to that of editing a word document. The main contrast with word processing is that the information is structured with respect to concepts pertaining to the assessment and framework domains, enabling automatic transformation of the item design into a first draft implemented version that can be passed to another stage of the item production process.

Distinguishing Authoring from Runtime and Management Platform Technologies

It has become common practice in the e-learning community to strictly separate the platform dependent components from the learning content and the tools used to design and execute that content. TBA is now starting to follow the same trend; however, practices inherited from paper-and-pencil assessment as well as the additional complexity that arises from psychometric constraints and models, sophisticated scoring and new advanced frameworks has somehow slowed down the adoption of this concept. In addition, the level of integration of IT experts and psychometricians in the community remains low. This often leads to an incomplete global or systemic vision on both sides, so that a significant number of technology-based assessments are implemented following a silo approach centred on the competency to be measured and including all the functionalities in a single closed application. Whenever the construct or framework changes or the types of items increase over the long run, this model is no longer viable. In contrast, the platform approach and the strict separation of test management and delivery layers, together with the strict separation of item runtime management and authoring, are the only scalable solution in high-diversity situations.

Items as Interactive Composite Hypermedia

In order to fully exploit the most recent advances in computer media technologies, one should be able to combine in an integrative manner various types of interactive media, to enable various types of user interactions and functionalities. In cases for which ubiquity—making assessment available everywhere—is a strong requirement, modern Web technologies must be seriously considered. Indeed, even if they still suffer from poorer performance and the lack of some advanced features that can be found in platform-dedicated tools, they nevertheless provide the sufficiently rich set of interaction features that one needs in most assessments. In addition, these technologies are readily available on a wide range of cost-effective hardware platforms, with cost-effective licenses, or even open source license. Moreover, Web technologies in general enable very diversified types of deployment across networks (their initial vocation), as well as locally, on laptops or other devices. This important characteristic makes deployments very cost-effective and customizable in assessment contexts.

This notion dramatically changes the vision one may have about item authoring tools. Indeed, on one hand, IT developers build many current complex and interactive items through ground-breaking programming, while on the other hand very simple items with basic interactions and data collection modes, such as multiple-choice items, are most often built using templates or simple descriptive languages accessible to non-programmers (such as basic HTML).

There are currently no easy and user-friendly intermediate techniques between these two extremes. Yet, most often, and especially when items are built on according to dynamic stepwise scenarios, the system needs to define and control a series of behaviours and user interactions for each item. If we distance ourselves from the media *per se* (the image, the video, a piece of an animation or a sound file, for instance), we realize that a large deal of user interactions and system responses can be modelled as changes of state driven by events and messages triggered by the user and transmitted between item objects.

The role of the item developer is to instantiate the framework as a scenario and to translate this scenario into a series of content and testee actions. In paper-and-pencil assessments, expected testee actions are reified in the form of instructions, and the data collection consists uniquely in collecting an input from the test taker. Since a paper instrument cannot change its state during the assessment, no behaviour or response to the user can be embedded in it.

One of the fundamental improvements brought by technology to assessment is the capacity to embed system responses and behaviours into an instrument, enabling it to change its state in response to the test taker's manipulations. This means that in the instantiation of the framework in a technology-based assessment setting, the reification of expected testee action is no longer in the form of instructions only, but also programmed into interaction patterns between the subject and the instrument. These can be designed in such a way that they steer the subject towards the expected sequence of actions. In the meantime, one can also collect the history of the user interaction as part of the input as well as the explicit information input by the test taker. As a consequence, depending on the framework, the richness of the item arises from both the type of media content and the user interaction patterns that drive the state of a whole item and all its components over time.

This clearly brings up different concerns from an authoring tool perspective. First, just as if they were manipulating tools to create paper-and-pencil items, item developers must create separately non-interactive (or loosely interactive) media content in the form of texts, images or sounds. Each of these media encapsulates its own set of functionalities and attributes. Second, they will define the structure of their items in terms of their logic flows (stimulus, tasks or questions, response collection and so on). Third, they will populate the items with the various media they need. And, fourth, they will set up the interaction scheme between the user and the media and between the different media.

Such a high-level Model-View-Controller architecture for item authoring tools, based on XML (Bray et al. 2006, 2008) and Web technologies, results in highly cost-effective authoring processes. They are claimed to foster *wider access to high quality visual interfaces and shorter authoring cycles for multi-disciplinary teams* (Chatty et al. 2004). It first lets item developers use their favourite authoring tools to design media content of various types instead of learning complex new environments and paradigms. In most cases, several of these tools are available as open-source software. In addition, the formats manipulated by them are often open standards available at no cost from the Web community. Then, considering the constant evolution of assessment domains, constructs, frameworks and, finally, instrument specifications, one should be able to extend rapidly and easily the scope of interactions and/or type of media that should be encapsulated into the item. With the content separated from the layout and the behavioural parts, the inclusion of new, sophisticated media into the item and in the user-system interaction patterns is made very easy and cost-effective. In the field of science, sophisticated media, such as molecular structure manipulators and viewers, such as Jmol (Herráez 2007; Willighagen and Howard 2007) and RasMol (Sayle and Milner-White 1995; Bernstein 2000), interactive mathematical tools dedicated to space geometry or other simulations can be connected to other parts of the item. Mathematic notations or 3D scenes described in X3D (Web3D Consortium 2007, 2008) or MathML (Carlisle et al. 2003) format, respectively, and authored with open-source tools, can also be embedded and connected into the interaction patterns of the items, together with SVG (Ferraiolo et al. 2009) images and XUL (Mozilla Foundation) or XAML (Microsoft) interface widgets, for instance. These principles have been implemented in the eXULiS package (Jadoul et al. 2006), as illustrated in Fig. 4.16. A conceptually similar but technically different approach, in which a conceptual model of an interactive media undergoes a series of transformations to produce the final executable, has been recently experimented with by Tissoires and Conversy (2008).

Going further along the transformational document approach, the document-oriented GUI enables users to directly edit documents on the Web, seeing that the Graphical User Interface is also a document (Draheim et al. 2006). Coupled with XML technologies and composite hypermedia item structure, this technique enables item authoring to be addressed as the editing of embedded layered documents describing different components or aspects of the item.

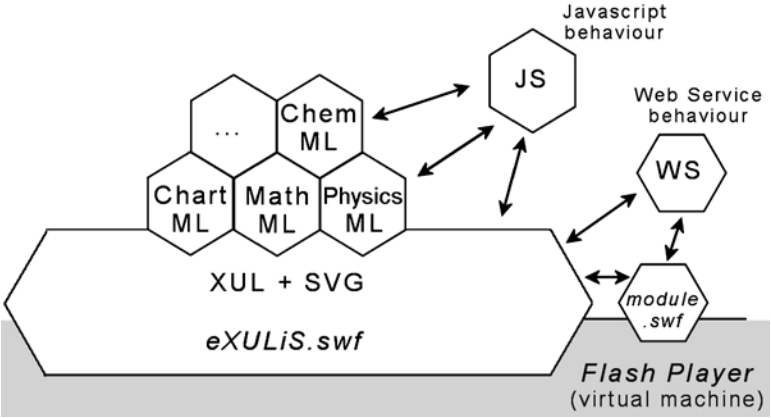


Fig. 4.16 Illustration of eXULiS handling and integrating different media types and services

Just as it has been claimed for the assessment resource management level, item authoring will also largely benefit from being viewed as a platform for interactive hypermedia integration. In a similar way as for the management platform, such a horizontal approach guarantees cost-effectiveness, time-effectiveness, openness and flexibility, while keeping the authoring complexity manageable.

Extending Item Functionalities with External On-Demand Services

The definitions of item behaviour and user interaction patterns presented above cover a large part of the item functional space. Composite interactive hypermedia can indeed accomplish most of the simple interactions that control the change of state of the item in response to user actions. However, there exist domains where more complex computations are expected at test time, during the test's administration. One can schematically distinguish four classes of such situations: when automatic feedback to the test taker is needed (mostly in formative assessments); when automatic scoring is expected for complex items; when using advanced theoretical foundations, such as Item Response Theory and adaptive testing and finally when the domain requires complex and very specific computation to drive the item's change of state, such as in scientific simulations.

When items are considered in a programmatic way, as a closed piece of software created by programmers, or when items are created from specialized software templates, these issues are dealt with at design or software implementation time, so that the complex computations are built-in functions of the items. It is very different when the item is considered as a composition of interactive hypermedia as was described above; such a built-in programmatic approach is no longer viable in the long run, the reasons being twofold. First, from the point of view of computational costs, the execution of these complex dedicated functions may be excessively time-consuming. If items are based on Web technologies and are client-oriented (the execution of the item functionalities is done on the client—the browser—rather than on the server), this may lead to problematic time lags between the act of the user and the computer's response. This is more than an ergonomic and user comfort issue; it may seriously endanger the quality of collected data. Second, from a cost and timeline point of view, proceeding in such a way implies lower reusability of components across domains and, subsequently, higher development costs, less flexibility, more iteration between the item developer and the programmer and, finally, longer delays.

Factorizing these functions out of the item framework constitutes an obvious solution. From a programmatic approach this would lead to the construction of libraries programmers can reuse for new items. In a more interesting, versatile and ubiquitous way, considering these functions as components that fits into the integrative composition of interactive hypermedia brings serious advantages. On one hand, it enables abstraction of the functions in the form of high-level software services that can be invoked by the item author (acting as an integrator of hypermedia and a designer of user-system interaction patterns) and on the other hand it enables higher reusability of components across domains. Moreover, in some circumstances, mostly depending on the deployment architecture, invocation of externalized software services may also partially solve the computational cost problem.

Once again, when looking at currently available and rapidly evolving technologies, Web technologies and service-oriented approaches, based on the UDDI (Clement et al. 2004), WSDL (Booth and Liu 2007) and SOAP (Gudgin et al. 2007) standards, offer an excellent ground for implementing this vision without drastic constraints on deployment modalities.

The added value of such an approach for externalizing software services can be illustrated in various ways. When looking at new upcoming frameworks and the general trend in education from content towards more participative inquiry-based learning, together with globalization and the increase of complexity of our modern societies, one expects that items will also follow the same transformations. Seeking to assess citizens' capacity to evolve in a more global and systemic multi-layered environment (as opposed to past local and strongly stratified environments where people only envision the nearby n +/− 1 levels) it seems obvious that constructs, derived frameworks and instantiated instruments and items will progressively take on the characteristics of globalized systems. This poses an important challenge for technology-based assessment that must support not only items and scenarios that are deterministic but also new ones that are not deterministic or are complex in nature. The complexity in this view is characterized either by a large response space when there exist many possible sub-optimal answers or by uncountable answers. This situation can typically occur in complex problem-solving where the task may refer to multiple concurrent objectives and yield to final solutions that may neither be unique nor consist of an optimum set of different sub-optimal solutions. Automatic scoring and, more importantly, management of system responses require sophisticated algorithms that must be executed at test-taking time. Embedding such algorithms into the item programming would increase dramatically the development time and cost of items, while lowering their reusability. Another source of complexity in this context that advocates the service approach arises when the interactive stimulus is a non-deterministic simulation (at the system level, not at local level of course). Multi-agent systems (often embedded in modern games) are such typical systems that are best externalized instead of being loaded onto the item.

In more classical instances, externalizing IRT algorithms in services invoked from the item at test-taking time will bring a high degree of flexibility for item designers and researchers. Indeed, various item models, global scoring algorithms and item selection strategies in adaptive testing can be tried out at low cost without modifying the core of existing items and tests. In addition, this enables the use of existing efficient packages instead of redeveloping the services. Another typical example can be found in science when one may need specific computation of energies or other quantities, or a particular simulation of a phenomenon. Once again, the service approach takes advantage of the existing efficient software that is available on the market. Last but not least, when assessing software, database or XML programming skills, some item designs include compilation or code execution feedbacks to the user at test-taking time. One would certainly never incorporate or develop a compiler or code validation into the item; the obvious solution rather is to call these tools as services (or Web services). This technique has been experimented in XML and SQL programming skill assessment in the framework of unemployed person training programme in Luxembourg (Jadoul and Mizohata 2006).

Finally, and to conclude this point, it seems that the integrative approach in item authoring is among the most scalable ones in terms of time, cost and item developer accessibility. Following this view, an item becomes a consistent composition of various interactive hypermedia and software services (whether interactive or not) that have been developed specifically for dedicated purposes and domains but are reusable across different situations rather than a closed piece of software or media produced from scratch for a single purpose. This reinforces the so-called horizontal platform approach to the cost of the current vertical full programmatic silo approach.

Item Banks, Storing Item Meta-data

Item banking is often considered to be the central element in the set of tools supporting computer-based assessment. Item banks are collections of items characterized by meta-data and most often collectively built by a community of item developers. Items in item banks are classified according to aspects such as difficulty, type of skill or topic (Conole and Waburton 2005).

A survey on item banks performed in 2004 reveals that most reviewed item banks had been implemented using SQL databases and XML technologies in various way; concerning meta-data, few had implemented meta-data beyond the immediate details of items (Cross 2004a). The two salient meta-data frameworks that arose from this study are derived from IEEE LOM (IEEE LTSC 2002) and IMS QTI

(IMS 2006). Since it is not our purpose here to discuss in detail the meta-data framework, but rather to discuss some important technologies that might support the management and use of semantically rich meta-data and item storage, the interested reader can refer to the IBIS report (Cross 2004b) for a more detailed discussion about meta-data in item banks.

When considering item storage, one should clearly separate the storage of the item *per se*, or its constituting parts, from the storage of meta-data. As already quoted by the IBIS report, relational databases remain today the favourite technology. However, with the dramatic uptake of XML-based technologies and considering the current convergence between the document approach and the interactive Web application approach around XML formats, the dedicated XML database can also be considered.

Computer-based assessment meta-data are used to characterize the different resources occurring in the various management processes, such as subjects and target groups, items and tests, deliveries and possibly results. In addition, in the item authoring process, meta-data can also be of great use in facilitating the search and exchange of media resources that will be incorporated into the items. This, of course, is of high importance when considering the integrative hypermedia approach.

As a general statement, meta-data can be used to facilitate

- Item retrieval when creating a test, concentrating on various aspects such as item content, purposes, models or other assessment qualities (the measurement perspective); the media content perspective (material embedded into the items); the construct perspective and finally the technical perspective (mostly for interoperability reasons)
- Correct use of items in consistent contexts from the construct perspective and the target population perspective
- Tracking usage history by taking into accounts the contexts of use, in relation to the results (scores, traces and logs)
- Extension of result exploitation by strengthening and enriching the link with diversified background information stored in the platform
- Sharing of content and subsequent economies of scale when inter-institutional collaborations are set up

Various approaches can be envisioned concerning the management of meta-data. Very often, meta-data are specified in the form of XML manifests that describe the items or other assessment resources. When exchanging, exporting or importing the resource, the manifest is serialized and transported together with the resource (sometimes the manifest is embedded into it). Depending on the technologies used to implement the item bank, these manifests are either stored as is, or parsed into the database. The later situation implies that the structure of the meta-data manifest is reflected into the database structure. This makes the implementation of the item bank dependent on the choice of a given meta-data framework, and moreover, that there is a common agreement in the community about the meta-data framework, which then constitutes an accepted standard. While highly powerful, valuable and generalized, with regard to the tremendous variability of assessment contexts and needs, one may rapidly experience the ‘standards curse’, the fact that there always exists a situation where the standard does not fit the particular need. In addition, even if this problem can be circumvented, interoperability issues may arise when one wishes to exchange resources with another system built according to another standard.

Starting from a fundamental stance regarding the need for a versatile and open platform as the only economically viable way to embrace assessment diversity and future evolution, a more flexible way to store and manage meta-data should be proposed in further platform implementation. Increasing the flexibility in meta-data management has two implications: first, the framework (or meta-data model, or meta-model) should be made updatable, and second, the data structure should be independent of the meta-data model. From an implementation point of view, the way the meta-data storage is organized and the way meta-data exploitation functions are implemented, this requires a soft-coding approach instead of traditional hard-coding. In order to do so, in a Web-based environment, Semantic Web (Berners-Lee et al. 2001) and ontology technologies are among the most promising technologies. As an example, such approach is under investigation for an e-learning platform to enable individual learners to use their own concepts instead of being forced to conform to a potentially inadequate standard (Tan et al. 2008). This enables one to annotate Learning Objects using ontologies (Gašević et al. 2004). In a more general stance, impacts and issues related to Semantic Web and ontologies in e-learning platforms have been studied by Vargas-Vera and Lytras (2008).

In the Semantic Web vision, Web resources are associated with the formal description of their semantics. The purpose of the semantic layer is to enable machine reasoning on the content of the Web, in addition to the human processing of documents. Web resource semantics is expressed as annotations of documents and services in meta-data that are themselves resources of the Web. The formalism used to annotate Web resources is triple model called the Resource Description Framework (RDF) (Klyne and Carrol 2004), serialized among other syntaxes in XML. The annotations make reference to a conceptual model called ontology and are modelled using the RDF Schema (RDFS) (Brickley and Guha 2004) or the Ontology Web language (OWL) (Patel-Schneider et al. 2004).

The philosophical notion of ontology has been extended in IT to denote the artefact produced after having studied the categories of things that exist or may exist in some domain. As such, the ontology results in a shared conceptualization of things that exist and make up the world or a subset of it, the domain of interest (Sowa 2000; Grubber 1993; Mahalingam and Huns 1997). An inherent characteristic of ontologies that makes them different from taxonomies is that they carry intrinsically the semantics of the concepts they describe (Grubber 1991; van der Vet and Mars 1998; Hendler 2001; Ram and Park 2004) with as many abstraction levels as required. Taxonomies present an external point of view of things, a convenient way to classify things according to a particular purpose. In a very different fashion, ontologies represent an internal point of view of things, trying to figure out how things are, as they are, using a representational vocabulary with formal definitions of the meaning of the terms together with a set of formal axioms that constrain the interpretation of these terms (Maedche and Staab 2001).

Fundamentally, in the IT field, ontology describes explicitly the structural part of a domain of knowledge within a knowledge-based system. In this context, 'explicit' means that there exists some language with precise primitives (Maedche and Staab 2001) and associated semantics that can be used as a framework for expressing the model (Decker et al. 2000). This ensures that ontology is machine processable and exchangeable between software and human agents (Guarino and Giaretta 1995; Cost et al. 2002). In some pragmatic situations, it simply consists of a formal expression of information units that describe meta-data (Khang and McLeod 1998).

Ontology-based annotation frameworks supported by RDF Knowledge-Based systems enable the management of many evolving meta-data frameworks with which conceptual structures are represented in the form of ontologies, together with the instances of these ontologies that represent the annotations. In addition, depending on the context, users can also define their own models in order to capture other features of assessment resources that are not considered in the meta-data framework. In the social sciences, such a framework is currently used to collaboratively build and discuss models on top of which surveys and assessments are built (Jadoul and Mizohata 2007).

Delivery Technologies

There is a range of methods for delivering computer-based assessments to students in schools and other educational institutions. The choice of delivery method needs to take account of the requirements of the assessment software, the computer resources in schools (numbers, co-location and capacity) and the bandwidth available for school connections to the Internet. Key requirements for delivery technologies are that they provide the basis for the assessment to be presented with integrity (uniformly and without delays in imaging), are efficient in the demands placed on resources and are effective in capturing student response data for subsequent analysis[4].

Factors Shaping Choice of Delivery Technology

The choice of delivery technology depends on several groups of factors. One of these is the nature of the assessment material; if it consists of a relatively simple stimulus material and multiple-choice response options to be answered by clicking on a radio button (or even has provision for a constructed text response) then the demands on the delivery technology will be relatively light. If the assessment includes rich graphical, video or audio material or involves students in using live software applications in an open authentic context then the demands on the delivery technology

[4] The contributions of Julian Fraillon of ACER and Mike Janic of SoNET systems to these thoughts are acknowledged.

will be much greater. For the assessment of twenty-first-century skills it is assumed that students would be expected to interact with relatively rich materials.

A second group of factors relates to the capacity of the connection of the school, or other assessment site, to the Internet. There is considerable variation among countries, and even among schools within countries, in the availability and speed of Internet connections in schools. In practice the capacity of the Internet connection needs to provide for simultaneous connection of the specified number of students completing the assessment, at the same time as other computer activity involving the Internet is occurring. There are examples where the demand of concurrent activity (which may have peaks) has not been taken into account. In the 2008 cycle of the Australian national assessment of ICT literacy, which involved ten students working concurrently with moderate levels of graphical material and interactive live software tasks but not video, a minimum of 4 Mbps was specified. In this project schools provided information about the computing resources and technical support that they had, by way of a project Web site that uses the same technology as the preferred test-delivery system so that the process of responding would provide information about Internet connectivity (and the capacity to use that connectivity) and the specifications of the computer resources available. School Internet connectivity has also proven to be difficult to monitor accurately. Speed and connectivity tests are only valid if they are conducted in the same context as the test taking. In reality it is difficult to guarantee this equivalence, as the connectivity context depends both on factors within schools (such as concurrent Internet and resource use across the school) and factors outside schools (such as competing Internet traffic from other locations). As a consequence it is necessary to cautiously overestimate the necessary connection speed to guarantee successful Internet assessment delivery. In the previously mentioned Australian national assessment of ICT literacy the minimum necessary standard of 4 Mbps per school was specified even though the assessment could run smoothly on a true connections speed of 1 Mbps.

A third group of factors relates to school computer resources, including sufficient numbers of co-located computers and whether those computers are networked. If processing is to be conducted on local machines it includes questions of adequate memory and graphic capacity. Whether processing is remote or local, screen size and screen resolution are important factors to be considered in determining an appropriate delivery technology. Depending on the software delivery solution being used it is also possible that school level software (in particular the type and version of the operating system and software plug-ins such as Java or ActiveX) can also influence the success of online assessment delivery.

Types of Delivery Technology

There is a number of ways in which computer-based assessments can be delivered to schools. These can be classified into four main categories: those that involve delivery through the Internet; those that work through a local server connected to the school network; those that involve delivery on removable media and those that involve delivery of mini-labs of computers to schools. The balance in the choice of delivery technology depends on a number of aspects of the IT context and changes over time, as infrastructure improves, existing technologies develop and new technologies emerge.

Internet-Based Delivery

Internet access to a remote server (typically using an SSL-VPN Internet connection to a central server farm) is often the preferred delivery method because the assessment software operates on the remote server (or server farm) and makes few demands on the resources of the school computers. Since the operation takes place on the server it provides a uniform assessment experience and enables student responses to be collected on the host server. This solution method minimizes, or even completely removes the need for any software installations on school computers or servers and eliminates the need for school technical support to be involved in setting up and execution. It is possible to have the remote server accessed using a thin client that works from a USB stick without any installation to local workstations or servers.

This delivery method requires a sufficient number of co-located networked computers with access to an Internet gateway at the school that has sufficient capacity for the students to interact with the material remotely without being compromised by other school Internet activity. The bandwidth required will depend on the nature of the assessment material and the number of students accessing it concurrently. In principle, where existing Internet connections are not adequate, it would be possible to provide school access to the Internet through a wireless network (e.g. Next G), but this is an expensive option for a large-scale assessment survey and is often least effective in remote areas where cable-based services are not adequate. In addition to requiring adequate bandwidth at the school, Internet-based delivery depends on the bandwidth and capacity of the remote server to accommodate multiple concurrent connections.

Security provisions installed on school and education system networks are also an issue for Internet delivery of computer-based assessments, as they can block access to some ports and restrict access to non-approved Internet sites. In general the connectivity of school Internet connections is improving and is likely to continue to improve; but security restrictions on school Internet access seem likely to become stricter. It is also often true that responsibility for individual school-level security rests with a number of different agencies. In cases where security is controlled at the school, sector and jurisdictional level the process of negotiating access for all schools in a representative large-scale sample can be extremely time consuming, expensive and potentially unsuccessful eventually.

A variant of having software located on a server is to have an Internet connection to a Web site but this usually means limiting the nature of the test materials to more static forms. Another variant is to make use of Web-based applications (such as Google docs) but this involves limitations on the scope for adapting those applications and on the control (and security) of collecting student responses. An advantage is that can provide the applications in many languages. A disadvantage is that if there is insufficient bandwidth in a school it will not be possible to locate the application on a local server brought to the school. In principle it would be possible to provide temporary connections to the Internet via the wireless network but at this stage this is expensive and not of sufficient capacity in remote areas.

Local Server Delivery

Where Internet delivery is not possible a computer-based assessment can be delivered on a laptop computer that has all components of the assessment software installed. This requires the laptop computer to be connected to the local area network (LAN) in the school and installed to operate (by running a batch file) as a local server with the school computers functioning as terminals. When the assessment is complete the student response data can delivered either manually (after being burned to CDs or memory sticks) or electronically (e.g. by uploading to an ftp site). The method requires a sufficient number of co-located networked computers and a laptop computer of moderate capacity to be brought to the school. This is a very effective delivery method that utilizes existing school computer resources but makes few demands on special arrangements.

Delivery on Removable Media

Early methods for delivering computer-based assessments to schools made use of compact disc (CD) technology. These methods of delivery limited the resources that could be included and involved complex provisions for capturing and delivering student response data. A variant that has been developed from experience of using laptop server technology is to deliver computer-based assessment software on Memory Sticks (USB or Thumb Drives) dispatched to schools by conventional means. The capacity of these devices is now such that the assessment software can work entirely from a Memory Stick on any computer with a USB interface. No software is installed on the local computer and the system can contain a database engine on the stick as well. This is a self-contained environment that can be used to securely run the assessments and capture the student responses. Data can then be delivered either manually (e.g. by mailing the memory sticks) or electronically (e.g. by uploading data to an ftp site). After the data are extracted the devices can be re-used. The pricing is such that even treating them as disposable is less than the cost of printing in a paper-based system. The method requires a sufficient number of co-located (but not necessarily networked) computers.

Provision of Mini-Labs of Computers

For schools with insufficient co-located computers it is possible to deliver computer-based assessments by providing a set of student notebooks (to function as terminals) and a higher specification notebook to act as the server for those machines

(MCEETYA 2007). This set of equipment is called a mini-lab. The experience of this is that cable connection in the mini-lab is preferable to a wireless network because it is less prone to interference from other extraneous transmissions in some environments. It is also preferable to operate a mini-lab with a server laptop and clients for both cost considerations and for more effective data management. The assessment software is located on the 'server' laptop and student responses are initially stored on it. Data are transmitted to a central server either electronically when an Internet connection is available or sent by mail on USB drives or CDs. Although this delivery method sounds expensive for a large project, equipment costs have reduced substantially over recent years and amount to a relatively small proportion of total costs. The difficulty with the method is managing the logistics of delivering equipment to schools and moving that equipment from school to school as required.

Use of Delivery Methods

All of these delivery technologies can provide a computer-based assessment that is experienced by the student in an identical way if the computer terminals at which the student works are similar. It is possible in a single study to utilize mixed delivery methods to make maximum use of the resources in each school. However, there are additional costs of development and licensing when multiple delivery methods are used. For any of the methods used in large-scale assessments (and especially those that are not Internet-based) it is preferable to have trained test administrators manage the assessment process or, at a minimum, to provide special training for school coordinators.

It was noted earlier in this section that the choice of delivery technology depends on the computing environment in schools and the optimum methods will change over time as infrastructure improves, existing technologies develop and new technologies emerge. In the Australian national assessment of ICT Literacy in 2005 (MCEETYA 2007) computer-based assessments were delivered by means of mini-labs of laptop computers (six per lab use in three sessions per day) transported to each of 520 schools. That ensured uniformity in delivery but involved a complex exercise in logistics. In the second cycle of the assessment in 2008 three delivery methods were used: Internet connection to a remote server, a laptop connected as a local server on the school network and mini-labs of computers. The most commonly used method was the connection of a laptop to the school network as a local server, which was adopted in approximately 68% of schools. Use of an Internet connection to a remote server was adopted in 18% of schools and the mini-lab method was adopted in approximately 14%. The use of an Internet connection to a remote server was more common in some education systems than others and in secondary compared to primary schools (the highest being 34% of the secondary schools in one State). Delivery by mini-lab was used in 20% of primary schools and nine per cent of secondary schools. In the next cycle the balance of use of delivery technologies will change and some new methods (such those based on memory sticks) will be available. Similarly the choice of delivery method will differ among countries and education systems, depending on the infrastructure in the schools, the education systems and, more widely, the countries.

Need for Further Research and Development

In this section, we first present some general issues and directions for further research and development. Three main topics will be discussed, which are more closely related to the technological aspects of assessment and add further topics to those elaborated in the previous parts of this chapter. Finally, a number of concrete research themes will be presented that could be turned into research projects in the near future. These themes are more closely associated with the issues elaborated in the previous sections and focus on specific problems.

General Issues and Directions for Further Research

Migration Strategies

Compared to other educational computer technologies, computer-based assessment bears additional constraints related to measurement quality, as already discussed. If the use of new technologies is being sought to widen the range of skills and competencies one can address or to improve the instrument in its various aspects, special care should be taken when increasing the technological complexity or the richness of the user experience to maintain the objective of an unbiased high-quality measurement. Looking at new opportunities offered by novel advanced technologies, one can follow two different approaches: either to consider technological opportunities as a generator of assessment opportunities or to carefully analyse assessment needs so as to derive technological requirements that are mapped onto available solutions or translated into new solution designs. At first sight, the former approach sounds more innovative than the latter, which seems more classical. However, both carry advantages and disadvantages that should be mitigated by the assessment context and the associated risks. The ‘technology opportunistic’ approach has major inherent strength, already discussed in this chapter, in offering a wide range of new potential instruments providing a complete assessment landscape. Besides this strength, it potentially opens the door to new time- and cost-effective measurable dimensions that have never been thought of before. As a drawback, it currently has tremendous needs for long and costly validations. Underestimating this will certainly lead to the uncontrolled use and proliferation of appealing but invalid assessment instruments. The latter approach is not neutral either. While appearing more conservative and probably more suitable for mid- and high-stakes contexts as well as for systemic studies, it also carries inherent drawbacks. Indeed, even if it guarantees the production of well-controlled instruments and developments in measurement setting, it may also lead to mid- and long-term time-consuming and costly operations that may hinder innovation by thinking ‘in the box’. Away from the platform approach, it may bring value by its capacity to address very complex assessment problems with dedicated solutions but with the risk that discrepancies between actual technology literacy of the target population and ‘old-fashion’ assessments will diminish the subject engagement—in other words and to paraphrase the US Web-Based Education Commission (cited in Bennett 2001), measuring today’s skills with yesterday’s technology.

In mid- and high-stakes individual assessments or systemic studies, willingness to accommodate innovation while maintaining the trend at no extra cost (in terms of production as well as logistics) may seem to be elusive at first sight. Certainly, in these assessment contexts, unless a totally new dimension or domain is defined, disruptive innovation would probably never arise and may not be sought at all. There is, however, a strong opportunity for academic interest in performing ambitious validation studies using frameworks and instruments built on new technologies. Taking into account the growing intricacy of psychometric and IT issues, there is no doubt that the most successful studies will be strongly inter-disciplinary. The intertwining of computer delivery issues, in terms of cost and software/hardware universality, with the maintenance of trends and comparability represents the major rationale that calls for inter-disciplinarity.

Security, Availability, Accessibility and Comparability

Security is of utmost importance in high-stakes testing. In addition to assessment reliability and credibility, security issues may also strongly affect the business of major actors in the fields. Security issues in computer-based assessment depend on the purposes and contexts of assessments, and on processes, and include a large range of issues.

The International Standard Institute has published a series of normative texts covering information security, known as the ISO 27000 family. Among these standards, ISO 27001 specifies requirements for information security management systems, ISO 27002 describes the Code of Practice for Information Security Management and ISO 27005 covers the topic of information security risk management.

In the ISO 27000 family, information security is defined according to three major aspects: the preservation of *confidentiality* (ensuring that information is accessible only to those authorized to have access), the preservation of information *integrity* (guaranteeing the accuracy and completeness of information and processing methods) and the preservation of information *availability* (ensuring that authorized users have access to information and associated assets when required). Security issues covered by the standards are of course not restricted to technical aspects.

They also consider organizational and more social aspects of security management. For instance, leaving a copy of an assessment on someone's desk induces risks at the level of confidentiality and maybe also at the level of availability. Social engineering is also another example of a non-technical security thread for password protection. These aspects are of equal importance in both paper-and-pencil and computer-based assessment.

The control of test-taker identity is classically achieved using various flavours of login/ID and password protection. This can be complemented by additional physical ID verification. Proctoring techniques have also been implemented to enable test takers to start the assessment only after having checked if the right person is actually taking the test. Technical solutions making use of biometric identification may help to reduce the risks associated with identity. As a complementary tool, the generalization of electronic passports and electronic signatures should also be considered as a potential contribution to the improvement of identity control.

Traditionally, in high-stakes assessment, when the test is administered centrally, the test administrator is in charge of detecting and preventing cheating. A strict control of the subject with respect to assessment rules before the assessment takes place is a minimal requirement. Besides the control, a classical approach to prevent cheating is the randomization of items or the delivery of different sets of booklets with equal and proven difficulty. The latter solution should preferably be selected because randomization of items poses other fairness problems that might disadvantage or advantage some test takers (Marks and Cronje 2008). In addition to test administrator control, cheating detection can be accomplished by analysing the behaviour of the subject during test administration. Computer forensic principles have been applied to the computer-based assessment environment to detect infringement of assessment rules. The experiment showed that typical infringement, such as illegal communication making use of technology, use of forbidden software or devices, falsifying identity or gaining access to material belonging to another student can be detected by logging all computer actions (Laubscher et al. 2005).

Secrecy, availability and integrity of computerized tests and items, of personal data (to ensure privacy) and of the results (to prevent loss, corruption or falsifications) is usually ensured by classical IT solutions, such as firewalls at server level, encryptions, certificates and strict password policy at server, client and communication network levels, together with tailored organizational procedures.

Brain dumping is a severe problem that has currently not been circumvented satisfactorily in high-stakes testing. Brain dumping is a fraudulent practice consisting of participating in a high-stakes assessment session (paper-based or computer-based) in order to memorize a significant number of items. When organized at a sufficiently large scale with many fake test takers, it is possible to reconstitute an entire item bank. After having solved the items with domain experts, the item bank can be disclosed on the Internet or sold to assessment candidates. More pragmatically and in a more straightforward way, an entire item bank can also be stolen and further disclosed by simply shooting pictures of the screens using a mobile phone camera or miniaturized Webcams. From a research point of view, as well as from a business value point of view, this very challenging topic should be paid more attention by the research community. In centralized high-stakes testing, potential ways of addressing the brain dump problem and the screenshot problem are twofold. On one hand, one can evaluate technologies to monitor the test-taker activity on and around the computer and develop alert patterns, and on the other hand, one can design, implement and experiment with technological solutions at software and hardware levels to prevent test takers from taking pictures of the screen.

Availability of tests and items during the whole assessment period is also a crucial issue. In the case of Internet-based testing, various risks may be identified, such as hijacking of the Web site or denial of service attacks, among others. Considering the additional risks associated with cheating in general, the Internet is not yet suitable for high- or mid-stakes assessment. However, solutions might be found to make the required assessment and related technology available everywhere (ubiquitous) and at every time it is necessary while overcoming the technological divide.

Finally, we expect that, from a research and development perspective, the topic of security in high-stakes testing will be envisioned in a more global and multi-dimensional way, incorporating in a consistent solution framework for all the aspects that have been briefly described here.

Ensuring Framework and Instrument Compliance with Model-Driven Design

Current assessment frameworks tend to describe a subject area on two dimensions—the topics to be included and a range of actions that drive item difficulty. However, the frameworks do not necessarily include descriptions of the processes that subjects use in responding to the items. Measuring these processes depends on more fully described models that can then be used not only to develop the items or set of items associated with a simulation but also to determine the functionalities needed in the computer-based platform. The objective is to establish a direct link between the conceptual framework of competencies to be assessed and the structure and functionalities of the item type or template. Powerful modelling capacities can be exploited for that purpose, which would enable one to:

- Maintain the semantics of all item elements and interactions and to guarantee that any one of these elements is directly associated with a concept specified in the framework
- Maintain the consistency of the scoring across all sets of items (considering automatic, semi-automatic or human scoring)
- Help to ensure that what is measured is, indeed, what is intended to be measured
- Significantly enrich the results for advanced analysis by linking with complete traceability the performance/ability measurement, the behavioural/temporal data and the assessment framework

It is, however, important to note that while IT can offer a wide range of rich interactions that might be able to assess more complex or more realistic situations, IT may also entail other important biases if not properly grounded on a firm conceptual basis. Indeed, offering respondents interaction patterns and stimuli that are not part of a desired conceptual framework may introduce performance variables that are not pertinent to the measured dimension. As a consequence, realism and attractiveness, although they may add to motivation and playability, might introduce unwanted distortions to the measurement instead of enriching or improving it. To exploit the capabilities offered by IT for building complex and rich items and tests so as to better assess competencies in various domains, one must be able to maintain a stable, consistent and reproducible set of instruments. If full traceability between the framework and each instrument is not strictly maintained, the risk of mismatch becomes significantly higher, undermining the instrument validity and consequently the measurement validity. In a general sense, the chain of decision traceability in assessment design covers an important series of steps, from the definition of the construct, skill, domain or competency to the final refinement of computerized items and tests by way of the design of the framework, the design of items, the item implementation and the item production. At each step, the design and implementation have the greatest probability of improving quality if they refer to a clear and well-formed meta-model while systematically referring back to pieces from the previous steps.

This claim is at the heart of the Model-Driven Architecture (MDA) software design methodology proposed by the Object Management Group (OMG). Quality and interoperability arise from the independence of the system specification with respect to system implementation technology. The final system implementation in a given technology results from formal mappings of system design to many possible platforms (Poole 2001). In OMG's vision, MDA enables improved maintainability of software (consequently, decreased costs and reduced delays), among other benefits, breaking the myth of stand-alone application that they require in never-ending corrective and evolutionary maintenance (Miller and Mukerji 2003).

In a more general fashion, the approach relates to Model-Driven Engineering, which relies on a series of components. Domain-specific modelling languages (DSLM) are formalized using meta-models, which define the semantics and constraints of concepts pertaining to a domain and their relationships. These DSLM components are used by designers to express their design intention declaratively as instances of the meta-model within closed, common and explicit semantics (Schmidt 2006). Many more meta-models than the actual facets of the domain require can be used to embrace the complexity and to address specific aspects of the design using the semantics, paradigms and vocabulary of different experts specialized in each individual facet. The second fundamental component consists of transformation rules, engines and generators, which are used to translate the conceptual declarative design into another model closer to the executable system. This transformational pathway from the design to the executable system can include more than one step, depending on the number of aspects of the domain together with operational and organizational production processes. In addition to the abovementioned advantages in terms of interoperability, system evolution and maintenance, this separation of concerns has several advantages from a purely conceptual design point of view: First, it keeps the complexity at a manageable level; second, it segments design activities centred on each specialist field of expertise; third, it enables full traceability of design decisions.

The latter advantage is at the heart of design and final implementation quality and risk mitigation. As an example, these principles have been successfully applied in the fields of business process engineering to derive business processes and e-business transactions through model chaining by deriving economically meaningful business processes from value models obtained by transforming an initial business model (Bergholtz et al. 2005; Schmitt and Grégoire 2006). In the field of information systems engineering, Turki et al. have proposed an ontology-based framework to design an information system by means of a stack of models that address different abstractions of the problem as well as various facets of the domain, including legal constraints. Applying a MDE approach, their framework consists of a *conceptual map to represent ontologies* as well as *a set of mapping guidelines from conceptual maps into other object specification formalisms* (Turki et al. 2004). A similar approach has been used to transform natural language mathematical documents into computerized narrative structure that can be further manipulated (Kamareddine et al. 2007). That transformation relies on a chain of model instantiations that address different aspects of the document, including syntax, semantics and rhetoric (Kamareddine et al. 2007a, b).

The hypothesis and expectation is that such a design approach will ensure compliance between assessment intentions and the data collection instrument. Compliance is to be understood here as the ability to maintain the links between originating design concepts, articulated according to the different facets of the problem and derived artefacts (solutions), along all the steps of the design and production process. Optimizing the production process, reducing the cost by relying on (semi-) automatic model transformation between successive steps, enabling conceptual comparability of instruments and possibly measuring their equivalence or divergence, and finally the guarantee of better data quality with reduced bias, are among the other salient expected benefits.

The claim for a platform approach independent from the content, based on a knowledge modelling paradigm (including ontology-based meta-data management), has a direct relationship in terms of solution opportunities to tackling the challenge of formal design and compliance. Together with Web technologies enabling distant collaborative work through the Internet, one can envision a strongly promising answer to the challenges.

To set up a new assessment design framework according to the MDE approach, several steps should be taken, each requiring intensive research and development work. First, one has to identify the various facets of domain expertise that are involved in assessment design and organize them as an assessment design process. This step is probably the easiest one and mostly requires a process of formalization. The more conceptual spaces carry inherent challenges of capturing the knowledge and expertise of experts in an abstract way so as to build the reference meta-models and their abstract relationships, which will then serve as a basis to construct the specific model instances pertaining to each given assessment in all its important aspects. Once these models are obtained, a dedicated instrument design and production chain can be set up, and the process started. The resulting instances of this layer will consist of a particular construct, framework and item, depending on the facet being considered. Validation strategies are still to be defined, as well as design of support tools.

The main success factor of the operation resides fundamentally in interdisciplinarity. Indeed, to reach an adequate level of formalism and to provide the adequate IT support tools to designers, assessment experts should work in close collaboration with computer-based assessment and IT experts who can bring their well-established arsenal of more formal modelling techniques. It is expected that this approach will improve measurement quality by providing more formal definitions of the conceptual chain that links the construct concepts to the final computerized instrument, minimizing the presence of item features or content that bear little or no relationship to the construct. When looking at the framework facets, the identification of indicators and their relationships, the quantifiers (along with their associated quantities) and qualifiers (along with their associated classes), and the data receptors that enable the collection of information used to value or qualify the indicators, must all be unambiguously related to both construct definition and item interaction patterns. In addition, they must provide explicit and sound guidelines for item designers with regard to scenario and item characteristic descriptions. Similarly, the framework design also serves as a foundation from which to derive exhaustive and unambiguous requirements for the software adaptation of extension from the perspective of item interaction and item runtime software behaviour. As a next step, depending on the particular assessment characteristics, item developers will enrich the design by instantiating the framework in the form of a semantically embedded scenario, which includes the definition of stimulus material, tasks to be completed and response collection modes. Dynamic aspects of the items may also be designed in the form of storyboards. Taking into account the scoring rules defined in the framework, expected response patterns are defined. As a possible following step, IT specialists will translate the item design into a machine-readable item description format. This amounts to the transposition of the item from a conceptual design to a formal description of the design in computer form, transforming a *descriptive* version to an *executable* or *rendered* version. Following the integrative hypermedia approach, the models involved in this transformation are the various media models and the integrative model.

Potential Themes for Research Projects

This section presents a list of research themes. The themes listed here are not yet elaborated in detail. Some of them are closely related and highlight different aspects of the same issue. These questions may later be grouped and organized into larger themes, depending on the time frame, size and complexity of the proposed research project. Several topics proposed here may be combined with the themes proposed by other working groups to form larger research projects.

Research on Enhancing Assessment

Media Effect and Validity Issues

A general theme for further research is the comparability of results of traditional paper-based testing and of technology-based assessment. This question may be especially relevant when comparison is one of the main aspects of the assessment, e.g. when trends are established, or in longitudinal research when personal developmental trajectories are studied. What kinds of data collection strategies would help linking in such cases?

A further research theme is the correspondence between assessment frameworks and the actual items presented in the process of computerized testing. Based on the information identified in points 1–4, new methods can be devised to check this correspondence.

A more general issue is the transfer of knowledge and skills measured by technology. How far do skills demonstrated in specific technology-rich environment transfer to other areas, contexts and situations, where the same technology is not present? How do skills assessed in simulated environments transfer to real-life situations? (See Baker et al. 2008 for further discussion.)

Logging, Log Analysis and Process Mining

Particularly challenging is making sense of the hundreds of pieces of information students may produce when engaging in a complex assessment, such as a simulation. How to determine which actions are meaningful, and how to combine those pieces into evidence of proficiency, is an area that needs concentrated research. The work on evidence-centred design by Mislevy and colleagues represents one promising approach to the problem.

Included in the above lines but probably requiring special mention is the issue of response latency. In some tasks and contexts, timing information may have meaning for purposes of judging automaticity, fluency or motivation, whereas in other tasks or contexts, it may be meaningless. Determining in what types of tasks and contexts response latency might produce meaningful information needs research, including whether such information is more meaningful for formative than summative contexts.

Saving and Analysing Information Products

One of the possibilities offered by computer-based assessment is for students to be able to save information products for scoring/rating/grading on multiple criteria. An area for research is to investigate how raters grade such complex information products. There is some understanding of how raters grade constructed responses in paper-based assessments, and information products can be regarded as complex constructed responses. A related development issue is whether it might be possible to score/rate information products using computer technology.

Computer-based assessment has made it possible to store and organize information products for grading, but, most of the time, human raters are required. Tasks involved in producing information products scale differently from single-task items. A related but further issue is investigating the dimensionality of computer-based assessment tasks.

Using Meta-information for Adaptive Testing and for Comparing Groups

It will be important to investigate how the information gathered by innovative technology-supported methods might be used to develop new types of adaptive testing in low-stakes, formative or diagnostic contexts. This could include investigating whether additional contextual information can be used to guide the processes of item selection.

In addition there are questions about whether there are interactions with demographic groups for measures, such as latency, individual collaborative skills, the collection of summative information from formative learning sessions or participation in complex assessments such that the meaning of the measures is different for one group versus another? More precisely, do such measures as latency, individual collaborative skills, summative information from formative sessions, etc., have the same meaning in different demographic groups? For example, latency may have a different meaning for males versus females of a particular country or culture because one group is habitually more careful than the other.

Connecting Data of Consecutive Assessments: Longitudinal and Accountability Issues

The analysis of longitudinal assessment data to build model(s) of developmental trajectories in twenty-first-century skills would be a long-term research project. Two of the questions to be addressed with these data are: What kind of design will facilitate the building of models of learners' developmental trajectories in the new learning outcome domains; and how can technology support collecting, storing and analysing longitudinal data?

Whether there exist conditions under which formative information can be used for summative purposes without corrupting the value of the formative assessments, students and teachers should know when they are being judged for consequential purposes. If selected classroom learning sessions are designated as 'live' for purposes of collecting summative information, does that reduce the effectiveness of the learning session or otherwise affect the behaviour of the student or teacher in important ways?

Automated Scoring and Self-Assessment

Automated scoring is an area of research and development with great potential for practice. On the one hand, a lot of research has been recently carried out on automated scoring (see Williamson et al. 2006a, b). On the other hand, in practice, real-time automated scoring is used mostly in specific testing situations or is restricted to certain simple item types. Further empirical research is needed, e.g. to devise multiple scoring systems and to determine which scoring methods are more broadly applicable and how different scoring methods work in different testing contexts.

Assessment tools for self-assessment versus external assessment are an area of investigation that could be fruitful. Assessment tools should also be an important resource to support learning. When the assessment is conducted by external agencies, it is supported by a team of assessment experts, especially in the case of high-stakes assessment, whether these are made on the basis of analysis of interaction data or information products (in which case, the assessment is often done through the use of rubrics). However, how such tools can be made accessible to teachers (and even students) for learning support through timely and appropriate feedback is important

Exploring Innovative Methods and New Domains of Assessment

New Ways for Data Capture: Computer Games, Edutainment and Cognitive Neuroscience Issues

Further information may be collected by applying specific additional instruments. Eye tracking is already routinely used in several psychological experiments and could be applied in TBA for a number of purposes as well. How and to what extent can one use screen gaze tracking methods to help computer-based training? A number of specific themes may be proposed. For example, eye tracking may help item development, as problematic elements in the presentation of an item can be identified in this way. Certain cognitive processes that students apply when solving problems can also be identified. Validity issues may be examined in this way as well.

How can computer games be used for assessment, especially for formative assessment? What is the role of assessment in games? Where is the overlap between ‘edutainment’ and assessment? How can technologies applied in computer games be transferred to assessment? How can we detect an addiction to games? How can we prevent game addictions?

How can the methods and research results of cognitive/educational neuroscience be used in computer-based assessments? For example, how and to what extent can a brain wave detector be used in measuring tiredness and level of concentration?

Person–Material Interaction Analysis

Further research is needed for devising general methods for the analysis of person–material interaction. Developing methods of analysing ‘trace data’ or ‘interaction data’ is important. Many research proposals comment that it must be possible to capture a great deal of information about student interactions with material, but there are few examples of systematic approaches to such data consolidation and analysis.

There are approaches used in communication engineering that are worth studying from the perspective of TBA as well; how might ways of traditionally analysing social science data be extended by using these innovative data collection technologies? Such simplified descriptive information (called fingerprints) from trace information (in this case, the detailed codes of video records of classrooms) was collected in the TIMSS Video study. The next step is to determine what characteristics of trace data are worth looking at because they are indications of the quality of student learning.

Assessing Group Outcomes and Social Network Analysis

Assessing group as opposed to individual outcomes is an important area for future research. Outcomes of collaboration do not only depend on the communication skills and social/personal skills of the persons involved, as Scardamalia and Bereiter have pointed out in the context of knowledge building as a focus of collaboration. Often, in real life, a team of knowledge workers working on the same project do not come from the same expertise background and do not possess the same set of skills, so they contribute in different ways to achieving the final outcome. Individuals also gain important learning through the process, but they probably learn different things as well, though there are overlaps, of course. How could group outcomes be measured, and what kinds of group outcomes would be important to measure?

How, and whether or not, to account for the contributions of the individual to collaborative activities poses significant challenges. Collaboration is an important individual skill, but an effective collaboration is, in some sense, best judged by the group's end result. In what types of collaborative technology-based tasks might we also be able to gather evidence of the contributions of individuals, and what might that evidence be?

How is the development of individual outcomes related to group outcomes, and how does this interact with learning task design? Traditionally in education, the learning outcomes expected of everyone at the basic education level are the same—these form the curriculum standards. Does group productivity require a basic set of core competences from everyone in the team? Answers to these two questions would have important implications for learning design in collaborative settings.

How can the environments in which collaborative skills are measured be standardized? Can one or all partners in a collaborative situation be replaced by 'virtual' partners? Can collaborative activities, contexts and partners be simulated? Can collaborative skills be measured in a virtual group where tested individuals face standardized collaboration-like challenges?

Social network analysis, as well as investigating the way people interact with each other when they jointly work on a computerbased task, are areas demanding further work. In network-based collaborative work, interactions may be logged, e.g. recording with whom students interact when seeking help and how these interactions are related to learning. Network analysis software may be used to investigate the interactions among people working on computer-based tasks, and this could provide insights into collaboration. The methods of social network analysis have developed significantly in recent years and can be used to process large numbers of interactions.

Affective Issues

Affective aspects of CBA deserve systematic research. It is often assumed that people uniformly enjoy learning in rich technology environments, but there is evidence that some people prefer to learn using static stimulus material. The research issue would not just be about person–environment fit but would examine how interest changes as people work through tasks in different assessment environments.

Measuring emotions is an important potential application of CBA. How and to what extent can Webcam-based emotion detection be applied? How can information gathered by such instruments be used in item development? How can measurement of emotions be used in relation to measurement of other domains or constructs, e.g. collaborative or social skills?

Measuring affective outcomes is a related area that could be the focus of research. Should more general affective outcomes, such as ethical behaviour in cyberspace, be included in the assessment? If so, how can this be done?

References

ACT. *COMPASS*. http://www.act.org/compass/

Ainley, M. (2006). Connecting with learning: Motivation, affect and cognition in interest processes. *Educational Psychology Review, 18*(4), 391–405.

Ainley, J., Eveleigh, F., Freeman, C., & O'Malley, K. (2009). *ICT in the teaching of science and mathematics in year 8 in Australia: A report from the SITES survey*. Canberra: Department of Education, Employment and Workplace Relations.

American Psychological Association (APA). (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: American Psychological Association.

Anderson, R., & Ainley, J. (2010). Technology and learning: Access in schools around the world. In B. McGaw, E. Baker, & P. Peterson (Eds.), *International encyclopedia of education* (3rd ed.). Amsterdam: Elsevier.

Baker, E. L., Niemi, D., & Chung, G. K. W. K. (2008). Simulations and the transfer of problem-solving knowledge and skills. In E. Baker, J. Dickerson, W. Wulfeck, & H. F. O'Niel (Eds.), *Assessment of problem solving using simulations* (pp. 1–17). New York: Lawrence Erlbaum Associates.

Ball, S., et al. (2006). Accessibility in e-assessment guidelines final report. Commissioned by TechDis for the E-Assessment Group and Accessible E-Assessment. Report prepared by Edexcel. August 8, 2011. Available: http://escholarship.bc.edu/ojs/index.php/jtla/article/view/1663

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning and Assessment, 2*(3). August 8, 2011. Available: http://escholarship.bc.edu/ojs/index.php/jtla/article/view/1663

Bennett, R. E. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives, 9*(5). Available: http://epaa.asu.edu/epaa/v9n5.html

Bennett, R. E. (2006). Moving the field forward: Some thoughts on validity and automated scoring. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 403–412). Mahwah: Erlbaum.

Bennett, R. (2007, September). New item types for computer-based tests. Presentation given at the seminar, What is new in assessment land 2007, National Examinations Center, Tbilisi. Retrieved January 19, 2011, from http://www.naec.ge/uploads/documents/2007-SEM_Randy-Bennett.pdf

Bennett, R. E. (2009). *A critical look at the meaning and basis of formative assessment (RM-09–06)*. Princeton: Educational Testing Service.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9–17.

Bennett, R. E., Morley, M., & Quardt, D. (1998). *Three response types for broadening the conception of mathematical problem solving in computerized-adaptive tests (RR-98–45)*. Princeton: Educational Testing Service.

Bennett, R. E., Goodman, M., Hessinger, J., Ligget, J., Marshall, G., Kahn, H., & Zack, J. (1999). Using multimedia in large-scale computer-based testing programs. *Computers in Human Behaviour, 15*, 283–294.

Bennett, R. E., Morley, M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement, 24*, 294–309.

Bennett, R. E., Jenkins, F., Persky, H., & Weiss, A. (2003). Assessing complex problem-solving performances. *Assessment in Education, 10*, 347–359.

Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). Problem solving in technology-rich environments: A report from the NAEP technology-based assessment project (NCES 2007–466). Washington, DC: National Center for Education Statistics, US Department of Education. Available: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B, Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment, 6*(9). Available: http://escholarship.bc.edu/jtla/vol6/9/

Bennett, R. E., Persky, H., Weiss, A., & Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. *Journal of Technology, Learning, and Assessment, 8*(8). Available: http://escholarship.bc.edu/jtla/vol8/8

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *Journal of Technology, Learning and Assessment, 6*(1). Available: http://escholarship.bc.edu/jtla/vol6/1/

Bergholtz, M., Grégoire, B., Johannesson, P., Schmitt, M., Wohed, P., & Zdravkovic, J. (2005). Integrated methodology for linking business and process models with risk mitigation. International Workshop on Requirements Engineering for Business Need and IT Alignment (REBNITA 2005), Paris, August 2005. http://efficient.citi.tudor.lu/cms/efficient/content.nsf/0/4A938852840437F2C12573950056F7A9/$file/Rebnita05.pdf

Berglund, A., Boag, S., Chamberlin, D., Fernández, M., Kay, M., Robie, J., & Siméon, J. (Eds.) (2007). XML Path Language (XPath) 2.0. W3C Recommendation 23 January 2007. http://www.w3.org/TR/2007/REC-xpath20–20070123/

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web: A new form of web that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American, 284*, 34–43.

Bernstein, H. (2000). Recent changes to RasMol, recombining the variants. *Trends in Biochemical Sciences (TIBS), 25*(9), 453–455.

Blech, C., & Funke, J. (2005). Dynamis review: An overview about applications of the dynamis approach in cognitive psychology. Bonn: Deutsches Institut für Erwachsenenbildung. Available: http://www.die-bonn.de/esprid/dokumente/doc-2005/blech05_01.pdf

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.), *Educational evaluation: New roles, new means.* The 63rd yearbook of the National Society for the Study of Education, part 2 (Vol. 69) (pp. 26–50). Chicago: University of Chicago Press.

Booth, D., & Liu, K. (Eds.) (2007). Web Services Description Language (WSDL) Version 2.0 Part 0: Primer. W3C Recommendation 26 June 2007. http://www.w3.org/TR/2007/REC-wsdl20-primer-20070626

Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment & Evaluation in Higher Education, 24*(4), 413–426.

Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E., & Yergeau, F., Cowan, J. (Eds.) (2006). *XML 1.1* (2nd ed.), W3C Recommendation 16 August 2006. http://www.w3.org/TR/2006/REC-xml11–20060816/

Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E., & Yergeau, F. (Eds.) (2008). *Extensible Markup Language (XML) 1.0* (5th ed.) W3C Recommendation 26 November 2008. http://www.w3.org/TR/2008/REC-xml-20081126/

Brickley, D., & Guha, R. (2004). RDF vocabulary description language 1.0: RDF Schema. W3C Recommandation. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/

Bridgeman, B. (2009). Experiences from large-scale computer-based testing in the USA. In F. Scheuermann, & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 39–44). Luxemburg: Office for Official Publications of the European Communities.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16*, 191–205.

Carlisle, D., Ion, P., Miner, R., & Poppelier, N. (Eds.) (2003). *Mathematical Markup Language (MathML) Version 2.0* (2nd ed.). W3C Recommendation 21 October 2003. http://www.w3.org/TR/2003/REC-MathML2–20031021/

Carnegie Learning. *Cognitive Tutors*. http://www.carnegielearning.com/products.cfm

Catts, R., & Lau, J. (2008). *Towards information literacy indicators*. Paris: UNESCO.

Chatty, S., Sire, S., Vinot J.-L., Lecoanet, P., Lemort, A., & Mertz, C. (2004). Revisiting visual interface programming: Creating GUI tools for designers and programmers. *Proceedings of UIST'04*, October 24–27, 2004, Santa Fe, NM, USA. ACM Digital Library.

Clement, L., Hately, A., von Riegen, C., & Rogers, T. (2004). *UDDI Version 3.0.2, UDDI Spec Technical Committee Draft, Dated 20041019.* Organization for the Advancement of Structured Information Standards (OASIS). http://uddi.org/pubs/uddi-v3.0.2–20041019.htm

Clyman, S. G., Melnick, D. E., & Clauser, B. E. (1995). Computer-based case simulations. In E. L. Mancall & P. G. Bashook (Eds.), *Assessing clinical reasoning: The oral examination and alternative methods* (pp. 139–149). Evanston: American Board of Medical Specialties.

College Board. *ACCUPLACER*. http://www.collegeboard.com/student/testing/accuplacer/

Conole, G., & Waburton, B. (2005). A review of computer-assisted assessment. *ALT-J, Research in Learning Technology, 13*(1), 17–31.

Corbiere, A. (2008). A framework to abstract the design practices of e-learning system projects. In *IFIP international federation for information processing*, Vol. 275; Open Source Development, Communities and Quality; Barbara Russo, Ernesto Damiani, Scott Hissam, Björn Lundell, Giancarlo Succi (pp. 317–323). Boston: Springer.

Cost, R., Finin, T., Joshi, A., Peng, Y., Nicholas, C., Soboroff, I., Chen, H., Kagal, L., Perich, F., Zou, Y., & Tolia, S. (2002). ITalks: A case study in the semantic web and DAML+OIL. *IEEE Intelligent Systems, 17*(1), 40–47.

Cross, R. (2004a). Review of item banks. In N. Sclater (Ed.), *Final report for the Item Bank Infrastructure Study (IBIS)* (pp. 17–34). Bristol: JISC.

Cross, R. (2004b). Metadata and searching. In N. Sclater (Ed.), *Final report for the Item Bank Infrastructure Study (IBIS)* (pp. 87–102). Bristol: JISC.

Csapó, B., Molnár, G., & R. Tóth, K. (2009). Comparing paper-and-pencil and online assessment of reasoning skills. A pilot study for introducing electronic testing in large-scale assessment in Hungary. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based* *assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 113–118). Luxemburg: Office for Official Publications of the European Communities.

CTB/McGraw-Hill. *Acuity*. http://www.ctb.com/products/product_summary.jsp?FOLDER%3C%3Efolder_id=1408474395292638

Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann, M., & Horrocks, I. (2000). The semantic web: The roles of XML and RDF. *IEEE Internet Computing, 15*(5), 2–13.

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In E. Spada & P. Reiman (Eds.), *Learning in humans and machine: Towards an interdisciplinary learning science* (pp. 189–211). Oxford: Elsevier.

Draheim, D., Lutteroth, C., & Weber G. (2006). Graphical user interface as documents. In *CHINZ 2006—Design Centred HCI*, July 6–7, 2006, Christchurch. ACM digital library.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Westport: American Council on Education/Praeger.

EDB (Education Bureau of the Hong Kong SAR Government) (2007). *Right Technology at the Right Time for the Right Task*. Author: Hong Kong.

Educational Testing Service (ETS). *Graduate Record Examinations (GRE)*. http://www.ets.org/portal/site/ets/menuitem.fab2360b1645a1de9b3a0779f1751509/?vgnextoid=b195e3b5f64f4010VgnVCM10000022f95190RCRD

Educational Testing Service (ETS). *Test of English as a foreign language iBT (TOEFL iBT)*. http://www.ets.org/portal/site/ets/menuitem.fab2360b1645a1de9b3a0779f1751509/?vgnextoid=69c0197a484f4010VgnVCM10000022f95190RCRD&WT.ac=Redirect_ets.org_toefl

Educational Testing Service (ETS). *TOEFL practice online*. http://toeflpractice.ets.org/

Eggen, T., & Straetmans, G. (2009). Computerised adaptive testing at the entrance of primary school teacher training college. In F. Sheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 134–144). Luxemburg: Office for Official Publications of the European Communities.

EMB (Education and Manpower Bureau HKSAR) (2001). Learning to learn – The way forward in curriculum. Retrieved September 11, 2009, from http://www.edb.gov.hk/index.aspx?langno=1&nodeID=2877

Ferraiolo, J., Jun, J., & Jackson, D. (2009). Scalable Vector Graphics (SVG) 1.1 specification. W3C Recommendation 14 January 2003, edited in place 30 April 2009. http://www.w3.org/TR/2003/REC-SVG11–20030114/

Feurzeig, W., & Roberts, N. (1999). *Modeling and simulation in science and mathematics education*. New York: Springer.

Flores, F.,Quint, V., & Vatton, I. (2006). Templates, microformats and structured editing. *Proceedings of DocEng'06, ACM Symposium on Document Engineering*, 10–13 October 2006 (pp. 188–197), Amsterdam, The Netherlands.

Gallagher, A., Bennett, R. E., Cahalan, C., & Rock, D. A. (2002). Validity and fairness in technology-based assessment: Detecting construct-irrelevant variance in an open-ended computerized mathematics task. *Educational Assessment, 8*, 27–41.

Gašević, D., Jovanović, J., & Devedžić, V. (2004). Ontologies for creating learning object content. In M. Gh. Negoita, et al. (Eds.), *KES 2004, LNAI 3213* (pp. 284–291). Berlin/Heidelberg: Springer.

Graduate Management Admission Council (GMAC). *Graduate Management Admission Test (GMAT)*. http://www.mba.com/mba/thegmat

Greiff, S., & Funke, J. (2008). Measuring complex problem solving: The MicroDYN approach. Heidelberg: Unpublished manuscript. Available: http://www.psychologie.uni-heidelberg.de/ae/allg/forschun/dfg_komp/Greiff&Funke_2008_MicroDYN.pdf

Grubber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*, 199–220.

Gruber, T. (1991 April). The role of common ontology in achieving sharable, reuseable knowledge bases. *Proceedings or the Second International Conference on Principles of Knowledge Representation and Reasoning* (pp. 601–602). Cambridge, MA: Morgan Kaufmann.

Guarino, N., & Giaretta, P. (1995). Ontologies and knowledge bases: Towards a terminological clarification. In N. Mars (Ed.), *Towards very large knowledge bases: Knowledge building and knowledge sharing* (pp. 25–32). Amsterdam: Ios Press.

Gudgin, M., Hadley, M., Mendelsohn, N., Moreau, J. -J., Nielsen, H., Karmarkar, A., & Lafon, Y. (Eds.) (2007). *SOAP Version 1.2 Part 1: Messaging framework* (2nd ed.). W3C Recommendation 27 April 2007. http://www.w3.org/TR/2007/REC-soap12-part1–20070427/

Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research, 17*(4), 397–431.

Hadwin, A., Winne, P., & Nesbit, J. (2005). Roles for software technologies in advancing research and theory in educational psychology. *The British Journal of Educational Psychology, 75*, 1–24.

Haldane, S. (2009). Delivery platforms for national and international computer based surveys. In F. Sheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 63–67). Luxemburg: Office for Official Publications of the European Communities.

Halldórsson, A., McKelvie, P., & Bjornsson, J. (2009). Are Icelandic boys really better on computerized tests than conventional ones: Interaction between gender test modality and test performance. In F. Sheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 178–193). Luxemburg: Office for Official Publications of the European Communities.

Hendler, J. (2001). Agents and the semantic web. *IEEE Intelligent Systems, 16*(2), 30–37.

Henri, F. (1992). Computer conferencing and content analysis. In A. R. Kaye (Ed.), *Collaborative learning through computer conferencing* (pp. 117–136). Berlin: Springer.

Herráez, A. (2007). *How to use Jmol to study and present molecular structures* (Vol. 1). Morrisville: Lulu Enterprises.

Horkay, N., Bennett, R. E., Allen, N., & Kaplan, B. (2005). Online assessment in writing. In B. Sandene, N. Horkay, R. E. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje (Eds.), Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project (NCES 2005–457). Washington, DC: National Center for Education Statistics, US Department of Education. Available: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment, 5*(2). Available: http://escholarship.bc.edu/jtla/vol5/2/

IEEE LTSC (2002). *1484.12.1-2002 IEEE Standard for Learning Object Metadata*. Computer Society/Learning Technology Standards Committee. http://www.ieeeltsc.org:8080/Plone/working-group/learning-object-metadata-working-group-12.

IMS (2006). IMS question and test interoperability overview, Version 2.0 Final specification. IMS Global Learning Consortium, Inc. Available: http://www.imsglobal.org/question/qti_v2p0/imsqti_oviewv2p0.html

International ICT Literacy Panel (Educational Testing Service). (2002). *Digital transformation: A framework for ICT literacy*. Princeton: Educational Testing Service.

Jadoul, R., & Mizohata, S. (2006). PRECODEM, an example of TAO in service of employment. *IADIS International Conference on Cognition and Exploratory Learning in Digital Age, CELDA 2006*, 8–10 December 2006, Barcelona. https://www.tao.lu/downloads/publications/CELDA2006_PRECODEM_paper.pdf

Jadoul, R., & Mizohata, S. (2007). Development of a platform dedicated to collaboration in the social sciences. Oral presentation at *IADIS International Conference on Cognition and Exploratory Learning in Digital Age, CELDA 2007*, 7–9 December 2007, Carvoeiro. https://www.tao.lu/downloads/publications/CELDA2007_Development_of_a_Platform_paper.pdf

Jadoul, R., Plichart, P., Swietlik, J., & Latour, T. (2006). eXULiS – a Rich Internet Application (RIA) framework used for eLearning and eTesting. *IV International Conference on Multimedia* *and Information and Communication Technologies in Education, m-ICTE 2006*. 22–25 November, 2006, Seville. In A. Méndez-Vilas, A. Solano Martin, J. Mesa González, J. A. Mesa González (Eds.), *Current developments in technology-assisted education*, Vol. 2. FORMATEX, Badajoz (2006), pp. 851–855. http://www.formatex.org/micte2006/book2.htm

Johnson, M., & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and Assessment, 4*(5), 311–326.

Kamareddine, F., Lamar, R., Maarek, M., & Wells, J. (2007). Restoring natural language as a computerized mathematics input method. In M. Kauers, et al. (Eds.), *MKM/Calculemus 2007, LNAI 4573* (pp. 280–295). Berlin/Heidelberg: Springer. http://dx.doi.org/10.1007/978–3–540–73086–6_23

Kamareddine, F., Maarek, M., Retel, K., & Wells, J. (2007). Narrative structure of mathematical texts. In M. Kauers, et al. (Eds.), *MKM/Calculemus 2007, LNAI 4573* (pp. 296–312). Berlin/Heidelberg: Springer. http://dx.doi.org/10.1007/978–3–540–73086–6_24

Kane, M. (2006). Validity. In R. L. Linn (Ed.), *Educational Measurement* (4th ed., pp. 17–64). New York: American Council on Education, Macmillan Publishing.

Kay, M. (Ed.) (2007). XSL Transformations (XSLT) Version 2.0. W3C Recommendation 23 January 2007. http://www.w3.org/TR/2007/REC-xslt20–20070123/

Kelley, M., & Haber, J. (2006). *National Educational Technology Standards for Students (NETS*S): Resources for assessment*. Eugene: The International Society for Technology and Education.

Kerski, J. (2003). The implementation and effectiveness of geographic information systems technology and methods in secondary education. *Journal of Geography, 102*(3), 128–137.

Khang, J., & McLeod, D. (1998). Dynamic classificational ontologies: Mediation of information sharing in cooperative federated database systems. In M. P. Papazoglou & G. Sohlageter (Eds.), *Cooperative information systems: Trends and direction* (pp. 179–203). San Diego: Academic.

Kia, E., Quint, V., & Vatton, I. (2008). XTiger language specification. Available: http://www.w3.org/Amaya/Templates/XTiger-spec.html

Kingston N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22*(1), 22–37.

Klyne, G., & Carrol, J. (2004). Resource description framework (RDF): Concepts and abstract syntax. W3C Recommendation. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

Koretz, D. (2008). *Measuring up. What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Kyllonen, P. (2009). New constructs, methods and directions for computer-based assessment. In F. Sheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 151–156). Luxemburg: Office for Official Publications of the European Communities.

Kyllonen, P., & Lee, S. (2005). Assessing problem solving in context. In O. Wilhelm & R. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 11–25). Thousand Oaks: Sage.

Latour, T., & Farcot, M. (2008). An open source and large-scale computer-based assessment platform: A real winner. In F. Scheuermann & A. Guimaraes Pereira (Eds.), *Towards a research agenda on computer-based assessment. Challenges and needs for European educational measurement* (pp. 64–67). Luxemburg: Office for Official Publications of the European Communities.

Laubscher, R., Olivier, M. S., Venter, H. S., Eloff, J. H., & Rabe, D. J. (2005). The role of key loggers in computer-based assessment forensics. In *Proceedings of the 2005 Annual Research Conference of the South African institute of Computer Scientists and information Technologists on IT Research in Developing Countries*, September 20–22, 2005,White River. SAICSIT (Vol. 150) (pp. 123–130). South African Institute for Computer Scientists and Information Technologists.

Lave, J. (1988). *Cognition in practice*. Cambridge: Cambridge University Press.

Law, N. (2005). Assessing learning outcomes in CSCL settings. In T.-W. Chan, T. Koschmann, & D. Suthers (Eds.), *Proceedings of the Computer Supported Collaborative Learning Conference (CSCL) 2005* (pp. 373–377). Taipei: Lawrence Erlbaum Associates.

Law, N., Yuen, H. K., Shum, M., & Lee, Y. (2007). *Phase (II) study on evaluating the effectiveness of the 'empowering learning and teaching with information technology' strategy (2004/2007). Final report*. Hong Kong: Hong Kong Education Bureau.

Law, N., Lee, Y., & Yuen, H. K. (2009). The impact of ICT in education policies on teacher practices and student outcomes in Hong Kong. In F. Scheuermann, & F. Pedro (Eds.), *Assessing the effects of ICT in education – Indicators, criteria and benchmarks for international comparisons* (pp. 143–164). Opoce: European Commission and OECD. http://bookshop.europa.eu/is-bin/INTERSHOP.enfinity/WFS/EU-Bookshop-Site/en_GB/-/EUR/ViewPublication-Start?PublicationKey=LB7809991

Lehtinen, E., Hakkarainen, K., Lipponen, L., Rahikainen, M., & Muukkonen, H. (1999). *Computer supported collaborative learning: A review. Computer supported collaborative learning in primary and secondary education.* A final report for the European Commission, Project, pp. 1–46.

Lie, H., & Bos, B. (2008). Cascading style sheets, level 1. W3C Recommendation 17 Dec 1996, revised 11 April 2008. http://www.w3.org/TR/2008/REC-CSS1–20080411

Linn, M., & Hsi, S. (1999). *Computers, teachers, peers: science learning partners*. Mahwah: Lawrence Erlbaum Associates.

Longley, P. (2005). *Geographic information systems and science*. New York: Wiley.

Lőrincz, A. (2008). Machine situation assessment and assistance: Prototype for severely handicapped children. In A. K. Varga, J. Vásárhelyi, & L. Samuelis (Eds.). In *Proceedings of Regional Conference on Embedded and Ambient Systems, Selected Papers* (pp. 61–68), Budapest: John von Neumann Computer Society. Available: http://nipg.inf.elte.hu/index.php?option=com_remository&Itemid=27&func=fileinfo&id=155

Macdonald, J. (2003). Assessing online collaborative learning: Process and product. *Computers in Education, 40*(4), 377–391.

Maedche, A., & Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems, 16*(2), 72–79.

Mahalingam, K., & Huns, M. (1997). An ontology tool for query formulation in an agent-based context. In *Proceedings of the Second IFCIS International Conference on Cooperative Information Systems,* pp. 170–178, June 1997, Kiawah Island, IEEE Computer Society.

Markauskaite, L. (2007). Exploring the structure of trainee teachers' ICT literacy: The main components of, and relationships between, general cognitive and technical capabilities. *Education Technology Research Development, 55*, 547–572.

Marks, A., & Cronje, J. (2008). Randomised items in computer-based tests: Russian roulette in assessment? *Journal of Educational Technology & Society, 11*(4), 41–50.

Martin, M., Mullis, I., & Foy, P. (2008). *TIMSS 2007 international science report. Findings from IEA's trends in international mathematics and science study at the fourth and eight grades.* Chestnut Hill: IEA TIMSS & PIRLS International Study Center.

Martin, R., Busana, G., & Latour, T. (2009). Vers une architecture de testing assisté par ordinateur pour l'évaluation des acquis scolaires dans les systèmes éducatifs orientés sur les résultats. In J.-G. Blais (Ed.), *Évaluation des apprentissages et technologies de l'information et de la communication, Enjeux, applications et modèles de mesure* (pp. 13–34). Quebec: Presses de l'Université Laval.

McConnell, D. (2002). The experience of collaborative assessment in e-learning. *Studies in Continuing Education, 24*(1), 73–92.

McDaniel, M., Hartman, N., Whetzel, D., & Grubb, W. (2007). Situational judgment tests: Response, instructions and validity: A meta-analysis. *Personnel Psychology, 60*, 63–91.

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers in Education, 39*(3), 299–312.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449–458.

Means, B., & Haertel, G. (2002). Technology supports for assessing science inquiry. In N. R. Council (Ed.), *Technology and assessment: Thinking ahead: Proceedings from a workshop* (pp. 12–25). Washington, DC: National Academy Press.

Means, B., Penuel, B., & Quellmalz, E. (2000). Developing assessments for tomorrowís classrooms. Paper presented at the The Secretary's Conference on Educational Technology 2000. Retrieved September 19, 2009, from http://tepserver.ucsd.edu/courses/tep203/fa05/b/articles/means.pdf

Mellar, H., Bliss, J., Boohan, R., Ogborn, J., & Tompsett, C. (Eds.). (1994). *Learning with artificial worlds: Computer based modelling in the curriculum*. London: The Falmer Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Microsoft. Extensible Application Markup Language (XAML). http://msdn.microsoft.com/en-us/library/ms747122.aspx

Miller, J., & Mukerji, J. (Eds.) (2003). MDA guide Version 1.0.1. Object Management Group. http://www.omg.org/cgi-bin/doc?omg/03–06–01.pdf

Ministerial Council for Education, Employment, Training and Youth Affairs (MCEETYA). (2007). *National assessment program – ICT literacy years 6 & 10 report*. Carlton: Curriculum Corporation.

Ministerial Council on Education, Early Childhood Development and Youth Affairs (MCEECDYA). (2008). *Melbourne declaration on education goals for young Australians*. Melbourne: Curriculum Corporation.

Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA). (1999). *National goals for schooling in the twenty first century*. Melbourne: Curriculum Corporation.

Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA). (2000). *Learning in an online world: The school education action plan for the information economy*. Adelaide: Education Network Australia.

Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA). (2005). *Contemporary learning: Learning in an on-line world*. Carlton: Curriculum Corporation.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centred design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). A brief introduction to evidence-centred design. (CSE Report 632). Los Angeles: UCLA CRESST.

Mislevy, R. J., Almond, R. G., Steinberg, L. S., & Lukas, J. F. (2006). Concepts, terminology, and basic models in evidence-centred design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15–47). Mahwah: Erlbaum.

Mozilla Foundation. *XML user interface language*. https://developer.mozilla.org/en/XUL_Reference

Mullis, I., Martin, M., Kennedy, A., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary school on 40 countries*. Chestnut Hill: Boston College.

Mullis, I., Martin, M., & Foy, P. (2008). *TIMSS 2007 international mathematics report. Findings from IEA's trends in international mathematics and science study at the fourth and eight grades.* Chestnut Hill: IEA TIMSS & PIRLS International Study Center.

Northwest Evaluation Association. *Measures of Academic Progress (MAP)*. http://www.nwea.org/products-services/computer-based-adaptive-assessments/map

OECD (2007). *PISA 2006 science competencies for tomorrow's world*. Paris: OECD.

OECD (2008a). Issues arising from the PISA 2009 field trial of the assessment of reading of electronic texts. Document of the 26th Meeting of the PISA Governing Board. Paris: OECD.

OECD (2008b). *The OECD Programme for the Assessment of Adult Competencies (PIAAC)*. Paris: OECD.

OECD (2009). *PISA CBAS analysis and results—Science performance on paper and pencil and electronic tests*. Paris: OECD.

OECD (2010). *PISA Computer-Based Assessment of Student Skills in Science*. Paris: OECD.

OMG. The object Management Group. http://www.omg.org/

Oregon Department of Education. *Oregon Assessment of Knowledge and Skills (OAKS)*. http://www.oaks.k12.or.us/resourcesGeneral.html

Patel-Schneider P., Hayes P., & Horrocks, I. (2004). OWL web ontology language semantics and abstract syntax. W3C Recommendation. http://www.w3.org/TR/2004/REC-owl-semantics-20040210/

Pea, R. (2002). *Learning science through collaborative visualization over the Internet.* Paper presented at the Nobel Symposium (NS 120), Stockholm.

Pearson. *PASeries*. http://education.pearsonassessments.com/pai/ea/products/paseries/paseries.htm

Pelgrum, W. (2008). School practices and conditions for pedagogy and ICT. In N. Law, W. Pelgrum, & T. Plomp (Eds.), *Pedagogy and ICT use in schools around the world: Findings from the IEA SITES 2006 study*. Hong Kong: CERC and Springer.

Pellegrino, J., Chudowosky, N., & Glaser, R. (2004). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Plichart P., Jadoul R., Vandenabeele L., & Latour T. (2004). TAO, a collective distributed computer-based assessment framework built on semantic web standards. In *Proceedings of the International Conference on Advances in Intelligent Systems—Theory and Application AISTA2004*, In cooperation with IEEE Computer Society, November 15–18, 2004, Luxembourg.

Plichart, P., Latour, T., Busana, G., & Martin, R. (2008). Computer based school system monitoring with feedback to teachers. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2008* (pp. 5065–5070). Chesapeake: AACE.

Plomp, T., Anderson, R. E., Law, N., & Quale, A. (Eds.). (2009). *Cross-national information and communication technology policy and practices in education* (2nd ed.). Greenwich: Information Age Publishing Inc.

Poggio, J., Glasnapp, D., Yang, X., & Poggio, A. (2004). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *Journal of Technology Learning, and Assessment, 3*(6), 30–38.

Poole, J. (2001). Model-driven architecture: Vision, standards and emerging technologies. Position paper in *Workshop on Metamodeling and Adaptive Object Models, ECOOP 2001*, Budapest, Hungary. Available: http://www.omg.org/mda/mda_files/Model-Driven_Architecture.pdf

Popper, K. (1972). *Objective knowledge: An evolutionary approach*. New York: Oxford University Press.

President's Committee of Advisors on Science and Technology, Panel on Educational Technology. (PCAST, 1997). *Report to the President on the use of technology to strengthen K-12 education in the United States*. Washington, DC: Author.

Quellmalz, E., & Haertel, G. (2004). Use of technology-supported tools for large-scale science assessment: Implications for assessment practice and policy at the state level: Committee on Test Design for K-12 Science Achievement. Washington, DC: Center for Education, National Research Council.

Quellmalz, E., & Pellegrino, J. (2009). Technology and testing. *Science, 323*(5910), 75.

Quellmalz, E., Timms, M., & Buckley, B. (2009). *Using science simulations to support powerful formative assessments of complex science learning*. Paper presented at the American Educational Research Association Annual Conference. Retrieved September 11, 2009, from http://simscientist.org/downloads/Quellmalz_Formative_Assessment.pdf

Raggett, D., Le Hors, A., & Jacobs, I. (1999). HTML 4.01 specification. W3C Recommendation 24 December 1999. http://www.w3.org/TR/1999/REC-html401–19991224

Ram, S., & Park, J. (2004). Semantic conflict resolution ontology (SCROL): An ontology for detecting and resolving data and schema-level semantic conflicts. *IEEE Transactions on Knowledge and Data Engineering, 16*(2), 189–202.

Reich, K., & Petter, C. (2009). eInclusion, eAccessibility and design for all issues in the context of European computer-based assessment. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 68–73). Luxemburg: Office for Official Publications of the European Communities.

Sakayauchi, M., Maruyama, H., & Watanabe, R. (2009). National policies and practices on ICT in education: Japan. In T. Plomp, R. E. Anderson, N. Law, & A. Quale (Eds.), *Cross-national information and communication technology policy and practices in education* (2nd ed., pp. 441–457). Greenwich: Information Age Publishing Inc.

Sandene, B., Bennett, R. E., Braswell, J., & Oranje, A. (2005). Online assessment in mathematics. In B. Sandene, N. Horkay, R. E. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje (Eds.), *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project* (NCES 2005–457). Washington, DC: National Center for Education Statistics, US Department of Education. Retrieved July 29, 2007 from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457

Sayle, R., & Milner-White, E. (1995). RasMol: Biomolecular graphics for all. *Trends in Biochemical Sciences (TIBS), 20*(9), 374.

Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.), *Liberal education in a knowledge society* (pp. 67–98). Chicago: Open Court.

Scardamalia, M., & Bereiter, C. (2003). Knowledge building environments: Extending the limits of the possible in education and knowledge work. In A. DiStefano, K. E. Rudestam, & R. Silverman (Eds.), *Encyclopedia of distributed learning* (pp. 269–272). Thousand Oaks: Sage.

Scheuermann, F., & Björnsson, J. (Eds.). (2009). *New approaches to skills assessment and implications for large-scale testing. The transition to computer-based assessment*. Luxembourg: Office for Official Publications of the European Communities.

Scheuermann, F., & Guimarães Pereira, A. (Eds.). (2008). *Towards a research agenda on computer-based assessment*. Luxembourg: Office for Official Publications of the European Communities.

Schmidt, D. C. (2006). Model-driven engineering. *IEEE Computer, 39*(2), 25–31.

Schmitt, M., & Grégoire, B., (2006). Business service network design: From business model to an integrated multi-partner business transaction. Joint International Workshop on Business Service Networks and Service oriented Solutions for Cooperative Organizations (BSN-SoS4CO '06), June 2006, San Francisco, California, USA. Available: http://efficient.citi.tudor.lu/cms/efficient/content.nsf/0/4A938852840437F2C12573950056F7A9/$file/Schmitt06_BusinessServiceNetworkDesign_SOS4CO06.pdf

Schulz, W., Fraillon, J., Ainley, J., Losito, B., & Kerr, D. (2008). *International civic and citizenship education study. Assessment framework*. Amsterdam: IEA.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Chicago: Rand McNally.

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher, 27*(2), 4.

Shermis, M. D., & Burstein, J. C. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah: Erlbaum.

Singapore Ministry of Education (1997). *Masterplan for IT in education: 1997–2002*. Retrieved August 17, 2009, from http://www.moe.gov.sg/edumall/mpite/index.html

Singleton, C. (2001). Computer-based assessment in education. *Educational and Child Psychology, 18*(3), 58–74.

Sowa, J. (2000). *Knowledge representation logical, philosophical, and computational foundataions*. Pacific-Groce: Brooks-Cole.

Stevens, R. H., & Casillas, A. C. (2006). Artificial neural networks. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 259–311). Mahwah: Erlbaum.

Stevens, R. H., Lopo, A. C., & Wang, P. (1996). Artificial neural networks can distinguish novice and expert strategies during complex problem solving. *Journal of the American Medical Informatics Association, 3*, 131–138.

Suchman, L. A. (1987). *Plans and situated actions. The problem of human machine communication*. Cambridge: Cambridge University Press.

Tan, W., Yang, F., Tang, A., Lin, S. & Zhang, X. (2008). An e-learning system engineering ontology model on the semantic web for integration and communication. In F. Li, et al. (Eds.). *ICWL 2008, LNCS 5145* (pp. 446–456). Berlin/Heidelberg: Springer.

Thompson, N., & Wiess, D. (2009). Computerised and adaptive testing in educational assessment. In F. Sheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 127–133). Luxemburg: Office for Official Publications of the European Communities.

Tinker, R., & Xie, Q. (2008). Applying computational science to education: The molecular workbench paradigm. *Computing in Science & Engineering, 10*(5), 24–27.

Tissoires, B., & Conversy, S. (2008). Graphic rendering as a compilation chain. In T. Graham, & P. Palanque (Eds.), *DSVIS 2008, LNCS 5136* (pp. 267–280). Berlin/Heidelberg: Springer.

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and education in twenty-eight countries: Civic knowledge and engagement at age fourteen*. Delft: IEA.

Turki, S., Aïdonis, Ch., Khadraoui, A., & Léonard, M. (2004). Towards ontology-driven institutional IS engineering. Open INTEROP Workshop on "*Enterprise Modelling and Ontologies for Interoperability*", *EMOI-INTEROP 2004*; Co-located with CaiSE'04 Conference, 7–8 June 2004, Riga (Latvia).

Van der Vet, P., & Mars, N. (1998). Bottom up construction of ontologies. *IEEE Transactions on Knowledge and Data Engineering, 10*(4), 513–526.

Vargas-Vera, M., & Lytras, M. (2008). Personalized learning using ontologies and semantic web technologies. In M.D. Lytras, et al. (Eds.). *WSKS 2008, LNAI 5288* (pp. 177–186). Berlin/Heidelberg: Springer.

Virginia Department of Education. *Standards of learning tests*. http://www.doe.virginia.gov/VDOE/Assessment/home.shtml#Standards_of_Learning_Tests

Wainer, H. (Ed.). (2000). *Computerised adaptive testing: A primer*. Hillsdale: Lawrence Erlbaum Associates.

Wang, S., Jiao, H., Young, M., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219–238.

Wang, S., Jiao, H., Young, M., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5–24.

Web3D Consortium (2007, 2008) ISO/IEC FDIS 19775:2008, Information technology—Computer graphics and image processing—Extensible 3D (X3D); ISO/IEC 19776:2007, Information technology—Computer graphics and image processing—Extensible 3D (X3D) encodings; ISO-IEC-19777–1-X3DLanguageBindings-ECMAScript & Java.

Webb, N. (1995). Group collaboration in assessment: Multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis, 17*(2), 239.

Weiss, D., & Kingsbury, G. (2004). Application of computer adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375.

Williamson, D. M., Almond, R. G., Mislevy, R. J., & Levy, R. (2006a). An application of Bayesian networks in automated scoring of computerized simulation tasks. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing*. Mahwah: Erlbaum.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006b). *Automated scoring of complex tasks in computer-based testing*. Mahwah: Erlbaum.

Willighagen, E., & Howard, M. (2007). Fast and scriptable molecular graphics in web browsers without Java3D. *Nature Precedings* 14 June. doi:10.1038/npre.2007.50.1. http://dx.doi.org/10.1038/npre.2007.50.1

Wirth, J., & Funke, J. (2005). Dynamisches Problemlösen: Entwicklung und Evaluation eines neuen Messverfahrens zum Steuern komplexer Systeme. In E. Klieme, D. Leutner, & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern* (pp. 55–72). Wiesbaden: VS Verlag für Sozialwissenschaften.

Wirth, J., & Klieme, E. (2003). Computer-based assessment of problem solving competence. *Assessment in Education: Principles, Policy & Practice, 10*(3), 329–345.

Xi, X., Higgins, D., Zechner, K., Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater v1.0* (RR-08–62). Princeton: Educational Testing Service.

Zhang, Y., Powers, D. E., Wright, W., & Morgan, R. (2003). *Applying the Online Scoring Network (OSN) to Advanced Placement program (AP) tests* (RM-03–12). Princeton: Educational Testing Service. Retrieved August 9, 2009 from http://www.ets.org/research/researcher/RR-03–12.html

Chapter 5
New Assessments and Environments for Knowledge Building

Marlene Scardamalia, John Bransford, Bob Kozma, and Edys Quellmalz

Abstract This chapter proposes a framework for integrating two different approaches to twenty-first century skills: "working backward from goals" and "emergence of new competencies." Working backward from goals has been the mainstay of educational assessment and objectives-based instruction. The other approach is based on the premise that breakthroughs in education to address twenty-first century needs require not only targeting recognized objectives but also enabling the discovery of new objectives—particularly capabilities and challenges that emerge from efforts to engage students in authentic knowledge creation. Accordingly, the focus of this chapter is on what are called "knowledge building environments." These are environments in which the core work is the production of new knowledge, artifacts, and ideas of value to the community—the same as in mature knowledge-creating organizations. They bring out things students are able to do that are obscured by current learning environments and assessments.

At the heart of this chapter is a set of developmental sequences leading from entry-level capabilities to the abilities that characterize members of high-performing knowledge-creating teams. These are based on findings from organization science and the learning sciences, including competencies that have already been demonstrated by students in knowledge-building environments. The same sources have been mined for principles of learning and development relevant to these progressions.

M. Scardamalia (✉)
University of Toronto, Canada
e-mail: marlene.scardamalia@utoronto.ca

J. Bransford
University of Washington, Seattle

B. Kozma
Kozmalone Consulting

E. Quellmalz
WestEd, San Francisco, California

P. Griffin et al. (eds.), *Assessment and Teaching of 21st Century Skills*,
DOI 10.1007/978-94-007-2324-5_5, © Springer Science+Business Media B.V. 2012

Knowledge Societies and the Need for Educational Reform

There is general agreement that the much-heralded "knowledge society" (Drucker 1994, 1968; Bell 1973; Toffler 1990) will have profound effects on educational, cultural, health, and financial institutions, and create an ever-increasing need for lifelong learning and innovation. This need for innovation is emphasized by the shift from manufacturing-based to knowledge-based economies, with the health and wealth of nations tied to the innovative capacity of its citizens and organizations. Furthermore, Thomas Homer-Dixon (2000) points out that problems such as global climate change, terrorism, information glut, antibiotic-resistant diseases, and the global financial crisis create an *ingenuity gap*: a critical gap between our need for ideas to solve complex problems and the actual supply of those ideas. More and more, prosperity—if not survival—will depend on innovation and the creation of new knowledge.

Citizens with little or poor education are particularly vulnerable. As David and Foray (2003) emphasize, disparities in productivity and growth of various countries have far less to do with their natural resources than with their capacity for creating new knowledge and ideas: "The 'need to innovate' is growing stronger as innovation comes closer to being the sole means to survive and prosper in highly competitive and globalized economies" (p. 22).

The call to action that launched this project, entitled *Transforming Education*: *Assessing and Teaching 21st Century Skills* (2009) stresses the need for systemic education reform to address the new challenges that confront us:

> The structure of global economy today looks very different than it did at the beginning of the 20th century, due in large part to advances in information and communications technologies (ICT). The economy of leading countries is now based more on the manufacture and delivery of information products and services than on the manufacture of material goods. Even many aspects of the manufacturing of material goods are strongly dependent on innovative uses of technologies. The start of the twenty-first century also has witnessed significant social trends in which people access, use, and create information and knowledge very differently than they did in previous decades, again due in many ways to the ubiquitous availability of ICT. These trends have significant implications for education. Yet most educational systems operate much as they did at the beginning of the 20th century and ICT use is far from ubiquitous. Significant reform is needed in education, world-wide, to respond to and shape global trends in support of both economic and social development (p.1).

According to one popular scenario, the introduction of technological advances into education will democratize knowledge and the opportunities associated with it. This may be too "romantic" a view, however. The current project is based on the assumption, shared by many (Laferrière 2001; Raizen 1997; Law 2006), that there is little reason to believe that technology combined with good intentions will be enough to make the kinds of changes that need to happen. To address these challenges, education reform must be systemic, not just technological. Systemic reform requires close ties between research-based innovation and practice (e.g., Bransford and Schwartz 2009), and assessment of progress, in order to create the know-how for knowledge-age education and workplace productivity. It also requires the alignment of organizational learning, policy, and the other components of the system (Bransford et al. 2000; Darling-Hammond 1997, 2000). As the call to action indicates:

> Systemic education reform is needed that includes curriculum, pedagogy, teacher training, and school organization. Reform is particularly needed in education assessment. … Existing models of assessment typically fail to measure the skills, knowledge, attitudes and characteristics of self-directed and collaborative learning that are increasingly important for our global economy and the fast-changing world (p.1).

Trilling and Fadel (2009) in their book *21st Century Skills: Learning for Life in Our Times* talk of "shifting-systems-in-sync." In order to judge different approaches to assessment, it is necessary to view them within the larger context of system dynamics in education. Traditionally, testing has played a part in a system that tends to stabilize at a level of mediocre performance and to be difficult to change. The system itself is well recognized and gives us such phenomena as the "mile wide, inch deep" curriculum, which no one advocates and yet which shows amazing persistence. Inputs to the system include standards, arrived at by consensus of educators and experts, tests geared to the standards, textbooks and other educational material geared to the standards and the tests, responses of learners to the curriculum (often manifested as failure to meet standards), responses of teachers, and pressures from parents (often focused on desire for their children to perform well on tests). These various elements interact until a state is reached that minimizes tensions between them. The typical result is standards that represent what tests are able to measure, teachers are comfortably able to teach, and students are comfortably able to learn. Efforts to introduce change may come from various sources, including new tests, but the system as a whole tends to nullify such efforts. This change-nullifying system has been well recognized by education leaders and has led to calls for "systemic reform." On balance, then, a traditional objectives- and test-driven approach is not a promising way to go about revolutionizing education or bringing it into the twenty-first century.

What are the alternatives? *How People Learn* (2000) and related publications from the National Academies Press have attempted to frame alternatives grounded in knowledge about brain, cognitive, and social development and embodying breakthrough results from experiments in the learning sciences. A rough summary of what sets these approaches apart from the one described above is elaborated below, including several examples that highlight the emergence of new competencies. In essence, instead of starting only with standards arrived at by consensus of stakeholders, these examples suggest the power of starting with what young learners are able to do under *optimal conditions (*Fischer & Bidell 1997; Vygotsky 1962/1934). The challenge then is to instantiate those conditions more widely, observe what new capabilities emerge, and work toward establishing conditions and environments that support "deep dives" into the curriculum (Fadel 2008). As the work proceeds, the goal is to create increasingly powerful environments to democratize student accomplishments and to keep the door open to further extensions of "the limits of the possible." This open-ended approach accordingly calls for assessments that are concurrent, embedded, and transformative, as we elaborate below. These assessments must be maximally useful to teachers and students so that they are empowered to achieve new heights. Formative assessment thus takes on a new meaning. It is integral to the learning process and connects communities (Earl 2003; Earl and Katz 2006). Instead of using it to narrow the gap between present performance and some targeted outcome, it is used to increase the distance between present performance and what has gone before, opening the door for exceeding targeted outcomes. It is additionally used to create increasingly effective knowledge-building environments that sustain such work and produce greater change over time.

In twenty-first century schools and other educational settings, knowledge and technological innovation will be inextricably related, as is currently the case in many knowledge-creating organizations, which provide models for high-level twenty-first century skills in action and the knowledge-building environments that support them. Once information and communication technology (ICT) becomes integral to the day-to-day, moment-to-moment workings of schools, organizations, and communities, a broad range of possibilities for extending and improving designs for knowledge-building environments and assessments follow. Accordingly, the goals for this chapter are to:

- Generate an analytic framework for analyzing environments and assessments that characterize and support knowledge-creating organizations and the knowledge-building environments that sustain them;
- Apply this framework to a set of environments and assessments in order to highlight models, possibilities, and variations in the extent to which they engage students in or prepare them for work in knowledge-creating organizations;
- Derive technological and methodological implications of assessment reform;
- Propose an approach to research that extends our understanding of knowledge-building environments and the needs and opportunities for promoting twenty-first century skills.

We start by discussing two concepts that underlie our whole treatment of assessment and teaching of twenty-first century skills: *knowledge-creating organizations* and *knowledge-building environments*.

Knowledge-Creating Organizations

A popular saying is that the future is here now; it's simply unevenly distributed. *Knowledge-creating organizations* are examples; they are companies, organizations, associations, and communities that have the creation, evaluation, and application of knowledge either as their main function or as an essential enabler of their main functions. Examples include research institutes, highly innovative companies, professional communities (medicine, architecture, law, etc.), design studios, and media production houses.

Creating new knowledge entails expectation and the means to go beyond current practice. Its goals are *emergent*, which means that they are formed and modified in the course of pursuing them. If computer design had not been characterized by emergent goals, computers would still be merely very fast calculating machines. Emergent outcomes cannot be traced back to subskills or subgoals, because they come about through self-organization—structure that arises from interactions among simpler elements that do not themselves foreshadow the structure. Color is a classic example of emergence; individual molecules do not have any color, but through self-organizing processes, molecular structures arise that do have color. System concepts are similarly applied to explaining the evolution of complex anatomical structures (Dawkins 1996) and to accounting for creativity (Simonton 1999)—one of the widely recognized twenty-first century skills. Creative work and adaptive expertise (Hatano and Inagaki 1986) alike are characterized by emergent goals. This makes them especially relevant to twenty-first century skills. The message here is not that "anything goes" and standards and visions should be abandoned. Instead, the message is that high standards and policies that support them must continually be "on the table" as something to be evaluated and exceeded, and that processes for innovation need to be supported, celebrated, assessed, and shared.

In a study by Barth (2009), "Over two-thirds of employers said that high school graduates were 'deficient' in problem solving and critical thinking." The importance of this point is highlighted by a survey in which about 3,000 graduates of the University of Washington, 5–10 years after graduation, rated the importance of various abilities they actually used in their work (Gillmore 1998). The top-ranked abilities were (1) *defining and solving problems*, (2) *locating information needed to help make decisions or solve problems*, (3) *working and/or learning independently*, (4) *speaking effectively*, and (5) *working effectively with modern technology, especially computers.* These were the abilities rated highest by graduates from all the major fields. Regardless of the students' field of study, these skills outranked knowledge and abilities specific to their field. They correspond fairly closely to items that appear on twenty-first century skill lists generated by business people and educators. Accordingly, it seems evident that they represent something important in contemporary work life, although precisely what they do represent is a question yet to be addressed.

The fact that so much of the pressure for teaching twenty-first century skills is coming from business people has naturally provoked some resistance among educators. Their main objections are to the effect that education should not be reduced to job training and that the private sector should not be dictating educational priorities. These are legitimate concerns, but they can be answered in straightforward ways:

- Teaching twenty-first century skills is a far cry from job training. It amounts to developing abilities believed to be of very broad application, not shaped to any particular kind of job. Indeed, as The North American Council for Online Learning and the Partnership for twenty-first Century Skills state (2006): "All citizens and workers in the twenty-first century must be able to think analytically and solve problems if they are to be successful—whether they are entry-level employees or high-level professionals" (p.7).

- Employability is an important consideration for today's students. Contrasting the changes taking place today with those of the Industrial Revolution, Peter Drucker (2003) has pointed out that very little relearning was required for a farm worker to become a factory worker, but that extensive learning and relearning is required for a factory worker to become a knowledge worker—learning that is best started in childhood.
- Crawford (2006) has questioned the emphasis on skills in the processing of abstract information. It is not expected that everyone will become what Reich (1991) called "symbolic analysts," but symbolic analysis and the use of technology for carrying it out are becoming increasingly essential for otherwise "manual" occupations (Leonard-Barton 1995).
- Well-accepted educational values require that whatever is done to promote twenty-first century skills should not be confined to the élite. It must be inclusive, foster equal participation, address issues of citizenship and multiculturalism, and provide for deliberative governance (Hearn and Rooney 2008; Robinson and Stern 1997; Trevinarus 1994, 2002).
- Increasing the level of knowledge-related skills is not only important for the managers and developers in an organization but also for empowering workers at all levels "to assume more responsibilities and solve problems themselves" (U.S. Department of Commerce et al. 1999, p.1).
- It is not assumed that modern corporations, research laboratories, design studios, and the like represent ideal models for education to emulate. There is probably as much to be learned from studying their shortcomings as from studying their successes. What they do represent, which is valuable for education systems, are social organizations that function to produce knowledge rather than merely to transfer and apply it. Thus they offer insight into a level of constructivism deeper than that characteristic of even the more active kinds of school learning (Scardamalia and Bereiter 2003).

The previous bullet point returns us to the theme of knowledge building and emergence. Instead of taking at face value the twenty-first century skills identified by committees of educators and business people, we might start by considering what constitutes knowledge creation at its best and what traits, abilities, and environments enable it. It is characteristic of "soft" skills of all kinds (of which twenty-first century skills are a subset) that everyone already possesses them to some degree (unlike "hard" skills, such as solving simultaneous equations and tooth filling, which may be totally lacking in the untrained). Thus for each skill identified as relevant to knowledge creation, we may establish a continuum running from the skill level almost everyone may be assumed to have, up to a level sufficient for engaging in creative knowledge work. The skills and competencies required for productive work in innovative organizations and professions provide a foundation for designing environments, practices, and formative assessments to help schools and education systems meet twenty-first century expectations (Trevinarus 1994, 2002; Wiggins and McTighe 2006; Anderson 2006).

Knowledge-Building Environments

The term "knowledge building" appears in approximately a million web documents. Sampling documents with a business orientation suggests the term is used as a synonym for "knowledge creation," roughly equivalent to concepts such as collective intelligence, intellectual capital, knowledge work, and innovation. Sampling education documents suggests it is used more as a synonym for "constructivist learning" (Wilson 1996), with rough equivalence to concepts such as active learning, discovery learning, and inquiry- and project-based learning.

The term "knowledge building" was originally introduced into the educational literature in 1989 (Bereiter and Scardamalia 1989, p.388) and had its basis in studies of expertise and innovation, summarized in the book *Surpassing Ourselves: An Inquiry Into the Nature and Implications of Expertise* (Bereiter and Scardamalia 1993). The phrase "progressive problem solving" was used to denote the process by which experts become experts and continue to develop their expertise (in contrast to becoming experienced nonexperts)—through investing their surplus cognitive resources in tackling problems at higher levels. The same basic idea, applied to knowledge building, took the form of a contrast between shallow and deep constructivism. If we imagine a line with shallow constructivism at one end and deep constructivism at the other, much of what is called "constructivist learning" in schools would be located toward the shallow end. Take for example the ubiquitous school "project" in which different project members assemble information that is then compiled in a multimedia presentation. One longtime observer of the school scene described this as using a computer to make a scrapbook. Knowledge building, with its focus on knowledge *creation*, would be located at the opposite end, aiming for the deepest levels of work with ideas, leading to emergence of new ideas and continued efforts to improve them (Scardamalia and Bereiter 2003).

Over the history of thought, the idea of knowledge as a human construction is relatively new. Designing environments to support knowledge creation is newer yet. Schools were not built for that purpose, and to this day many would claim they should not or could not. Yet universal access to the process whereby new knowledge is created arguably depends on bringing knowledge-building environments into schools.

In brief, we use the term "knowledge-building environments" to refer to contexts supportive of the emergence and further development of new ideas—knowledge creation in organizations of all kinds. Conceptually, economically, and technologically, it may be necessary to connect currently distinct environments for creative work with ideas (e.g., knowledgeware) to those for learning (e.g., courseware, tutorials, simulations), so as to encourage their integration and easy movement between these different and essential aspects of mature knowledge work. What would a more integrative approach look like? We say more about that below. For now, we elaborate the concept of knowledge-building environments by focusing on features that favor the emergence of new skills.

Knowledge-building environments provide special support for creative work with ideas, so that ideas may grow from nascent form to something of greater consequence than could have been imagined before. Improved ideas emerge as they are generated in multiple and varied contexts and are entered into communal spaces. Within these more public spaces, collaborators as well as competitors can elaborate, critique, reframe, link, re-position, create higher-order structures, explore and devise uses for ideas, and in other ways work creatively with them. It is through such sustained and varied engagement that ideas, like colorless molecules, acquire new properties through structural organization. In line with this *emergence* perspective, a knowledge-building approach considers the "promisingness" of an idea, recognizing that through new combinations and sustained work, something brilliant might emerge. In creative knowledge work it is important both to avoid wasting resources on unpromising ideas and to guard against killing off ideas that have promise. As the designer of a program for forest conservation remarked in response to criticisms of the plan, "an imperfect program which can be improved is better than none at all" ("Saving the rainforest: REDD or dead?" 2009).

In summary, a knowledge-building environment, virtual or otherwise, is one that enhances collaborative efforts to create and continually improve ideas. It exploits the potential of collaborative knowledge work by situating ideas in a communal workspace where others can criticize or contribute to their improvement. In these collaborative open contexts, discourse that is democratic and directed toward idea advancement compounds the value of ideas, so that collective achievement exceeds individual contributions. A local knowledge-building community gains strength as it connects to a broader one. The local community not only draws upon, but also affords participation in, the larger one, with possibilities for symmetrical advances of knowledge. A successful knowledge-building environment will bring innovation closer to the central work of an organization. It is an environment in which members are continually contributing to and enhancing the shared intellectual resources of the organization. Each advance precipitates another, so that at both the individual and group level, there is continual movement beyond current understanding and capacity. Emergence becomes a way of life, different from but both more productive and more personally satisfying than a life restricted to following known paths to known goals. Innovation, as Peter Drucker (1985, p.151) put it, becomes "part and parcel of the ordinary, the norm, if not routine."

New Goals and Methods to Support the Emergence of New Skills

Advocates for the adoption of twenty-first century skills generally look for this to have an overall transformative effect on the schools. However, the nature and extent of this envisaged transformation can range from conservative to fundamental, as suggested by the following three levels:

1. *Additive change*. Change is expected to result from the addition of new skill objectives, new curriculum content (nanotechnology, environmental studies,

cross-cultural studies, systems theory, technology studies, etc.), and new technology. Changes to existing curricula will be needed to make room for additions.

2. *Assimilative change*. Instead of treating work on twenty-first century skills as an add-on, existing curricula and teaching methods are modified to place greater emphasis on critical thinking, problem solving, collaboration, and so forth. This is the most widely recommended approach and reflects lessons learned from the disappointing results of a previous wave of "higher-order thinking skills" instruction that took the additive approach (Bereiter 1984).
3. *Systemic change*. Instead of incorporating new elements into a system that retains its nineteenth century structure, schools are transformed into twenty-first century organizations. Toward this end we present a case for schools to operate as knowledge-creating organizations. The envisaged educational change is not limited to schools, however. Knowledge creation by young people can and often does take place in out-of-school contexts.

The present authors clearly favor systemic change but recognize that the realities of public education often mean that assimilative change, and in many cases additive change, is as far as a school system will go in adapting to twenty-first century opportunities and needs. Accordingly, approaches to teaching and assessing twenty-first century skills need to be applicable and potentially transformative at any of the three levels. That said, however, we suggest that countries whose schools are transformed into knowledge-creating organizations may gain a tremendous advantage over those that struggle to incorporate knowledge-age education into industrial-age curricula and structures.

Two general strategies are applicable to pursuing the practical goals of advancing twenty-first century skills, and we argue that both are important and need to be used in a complementary fashion. One is the approach of *working backward from goals*. The other is one that, for reasons that will become evident, we call an *emergence* approach.

"Working backward from goals" to construct a system of subgoals and a path leading from an initial state to the goal is one of the main strategies identified in Newell and Simon's classic study of problem solving (1972). It will be recognized as the most frequently recommended way of designing instruction. As applied to educational assessment, it comprises a variety of techniques, all of which depend on a clearly formulated goal, the antecedents of which can be identified and separately tested. Although working backward is a strategy of demonstrable value in cases where goals are clear, it has two drawbacks in the case of twenty-first century skills. Most twenty-first century skills are "soft" skills, which means among other things that there is an inevitable vagueness and subjectivity in regard to goals, which therefore makes "working backward" not nearly so well structured as in the case of "hard" skills (such as the ability to execute particular algebraic operations). A more serious difficulty, however, is that working backward from goals provides no basis for discovering or inventing new goals—and if twenty-first century education is to be more than a tiresome replication of the 1970s "higher-order skills" movement, it has to be responsive to potential expansions of the range of what's possible.

As noted earlier, in the context of teaching and testing twenty-first century skills, "working backwards from goals" needs to be complemented by a working-forward approach growing out of what has been called the "systems revolution" (Ackoff 1974). Self-organization and emergence are key ideas in a systems approach to a vast range of problems. An "emergence" approach, when closely tied to educational experimentation, allows for the identification of new goals based on the discovered capabilities of learners. The observation that, in advance of any instruction in rational numbers, children possess an intuitive grasp of proportionality in some contexts led to formulation of a new goal (rational number sense) and development of a new teaching approach that reversed the traditional sequence of topics (Moss 2005). Results suggest that both the traditional goals (mastering appropriate algorithms) and the path to achieving them (starting by introducing rational numbers through models that connect children's whole number arithmetic) were misconceived, even though they were almost universally accepted. If that can happen even on such a well-traveled road as the teaching of arithmetic, we must consider how much riskier exclusive reliance on a working-backward approach might be to the largely untried teaching of twenty-first century skills. But the drawback of the emergence approach, of course, is that there is no guarantee that a path can be found to the emergent goal. Invention is required at every step, with all its attendant uncertainties.

Two concrete examples may help clarify the nature of an "emergence" approach and its benefits. The first example expands on the previously cited work of Moss (2005). The second example, drawn from work on scientific literacy, points to a potentially major twenty-first century skill that has gone unrecognized in the top-down and "working-backward" approaches that have dominated mainstream thinking about twenty-first century skills.

1. *Beyond rational number skills to proportional thinking*. Failure to master rational numbers is endemic and has been the subject of much research. Much of the difficulty, it appeared, is that students *transferred* their well-learned whole number arithmetic to fractions and thus failed to grasp the essential idea of proportionality, or the idea that fractions are numbers in their own right. The standard way of introducing fractions, via countable parts of a whole, was seen as reinforcing this tendency. Joan Moss and Robbie Case observed, however, that children already possessed an idea of proportionality, which they could demonstrate when asked to pour liquid into two different-sized beakers so that one was as full as the other. Once proportional reasoning was recognized as a realistic goal for mathematics teaching, "working backwards" could then be applied to devising ways of moving toward that goal. Moss (2005) developed a whole environment of artifacts and activities the purpose of which was to engage students in thinking proportionally. Instead of introducing fractions as the starting point for work on rational numbers, Moss and Case started with percentages, as being more closely related to spontaneous understanding (consider the bars on computer screens that register what percent of a task has been completed). In final assessments, students in grades 5 and 6 outperformed educated adults. Another name for proportional thinking is rational number sense. Greeno (1991) characterized number sense as knowing one's way around in a numerical domain, analogous to

knowing one's way around in a geographical area. It is not something that is directly taught but rather something that emerges from experience in crossing and recrossing a domain in different directions and with different purposes. It is assessable, but it is not specifiable in the way that hard skills are. And, quite obviously, proportional thinking or rational number sense is a more fundamental and more skill-enhancing outcome than mastering (or not quite mastering) a number of rational number algorithms.

2. *Beyond "scientific method" to theory building.* The second example of an emergence approach, more directly related to twenty-first century skills, comes from work on theory building. Broadly conceived, creative *knowledge* work of all kinds—planning, inventing, and so forth—is theory building. Even the Wright Brothers, known to the world as exceptionally clever tinkerers, were explicitly engaged in theory building at the same time they were engaged in building an airplane (Bereiter 2009). Ability to construct, test, and improve theory-like knowledge structures could therefore rate as a top-level twenty-first century skill. It does not appear on twenty-first century skill lists, however, possibly because it is not readily described in skill terms and because little is known about what students are capable of in this respect. Expert opinion has suggested that work on theory building should wait until high school (Smith and Wenk 2006) and that the learning progression should start with hypothesis testing and control of variables (Kuhn et al. 1992; Schauble et al. 1995). Instructional results from this approach have not been encouraging with respect to scientific literacy, and there have been many efforts to find new approaches (Carey et al. 1989; Carey and Smith 1993; Honda 1994; Smith et al. 2000), with further confirmation of the conventional expert wisdom that theory building is beyond the capacity of young students. When free to pursue problems of understanding on their own initiative, however, students were observed to engage spontaneously in a good deal of theorizing (Scardamalia and Bereiter 2006). A small experiment was carried out in which grade 4 students in a class where knowledge building was the norm were compared with similar students who had followed a more traditional inquiry approach (Bereiter and Scardamalia 2009). In the knowledge-building class there was no explicit teaching of "scientific method" and no carrying out of pre-specified experiments. Instead, the students were supported in creating, exploring, and considering theories from multiple perspectives. Results showed significantly higher levels of theoretical work and scientific literacy and superior scientific writing for the emergent goals approach (Bereiter and Scardamalia 2009; Chuy et al. 2009). Theory building, it turns out, is not only possible in 10- to 12-year-olds but also at even earlier ages. A kindergarten teacher in the same school learned of the findings and thought her students might have relevant, untapped capacities. She asked them to generate theories about why some trees in their schoolyard had no new leaves in the early spring while other trees did. The children not only generated a number of reasonable explanations but also connected these with supportive facts. It would seem, therefore, that theory building could justifiably gain a place among the twenty-first century skills to be developed and tested from early childhood onward. Work by Shutt et al. (2011) also supports this point of view.

In a later section, on technology for supporting the emergence of new competencies (pp. 237 ff.), we discuss the specific forms of support that have enabled the achievement of exceptional levels of proportional reasoning and theory development. As the preceding examples suggest, discovering new goals is not simply a matter of turning students loose in an environment and waiting to see what happens. Discovering new goals is an aspect of scientific discovery, and rarely is such discovery accidental. People know in a general way what they are looking for, and particular moves may be carefully calculated, but this process as a whole has to be structured so as to allow room for unexpected insights. When Darwin set sail on the beagle, he did not know he was about to explain the origin of species, but he was not merely a collector of curious specimens, either.

Most current school reform efforts, whether involving new management structures or the introduction of new standards and curricula, are additive as far as their treatment of twenty-first century skills is concerned. Changes are based on conservative practices and templates drawn from instruction in traditional subjects. More transformative change requires goals and methods to be considered anew. Education for twenty-first century skills may in fact have no "tried and true" methods to draw on, so riskier approaches are needed. It would be difficult to get excited about twenty-first century education reform were it nothing more than extending existing goals to more demanding performance levels. It should, of course, include such goals—performance demands are indeed likely to rise, and there will, no doubt, continue to be students who need help in meeting even today's modest standards. But anything that deserves the name of education for the twenty-first century needs new kinds of objectives, not simply higher standards for existing ones.

In the following sections, we examine twenty-first century skills as they are being enacted in knowledge-creating organizations. We focus on what is involved in the knowledge creation being carried out by experts actually working in these organizations, providing a sharpened focus for "working backward" to identify methods and goals that might apply to schools, while allowing us to go beyond the identification of the desirable traits and skills that are viewed by employers wishing to hire people for knowledge work. We then consider the knowledge-building environments that support work in knowledge-creating organizations, followed by examining learning and assessment theory. In the section on specific investigations, we propose investigations within an emergence framework, using findings from the working-backward approach to test transfer and generalization effects so as to achieve a best-of-both-worlds synthesis of working backward and emergence of new competencies.

Characteristics of Knowledge-Creating Organizations

How do businesses succeed in a knowledge economy? How are knowledge-intensive firms organized and how do they function? How are jobs different in a knowledge economy? And what kinds of skills are needed?

Industry- or firm-level studies in the USA (Stiroh 2003), the U.K. (Borghans and ter Weel 2001; Dickerson and Green 2004; Crespi and Pianta 2008), Canada (Gera and Gu 2004; Zohgi et al. 2007), France (Askenazy et al. 2001; Maurin and Thesmar 2004), Finland (Leiponen 2005), Japan (Nonaka and Takeuchi 1995), and Switzerland (Arvanitis 2005) have found many similar results—a major factor in the success of highly productive, innovative firms is the use of ICT (UNESCO 2005). Of course, productivity and innovation increases did not come merely with the introduction of new technologies. Rather, technology use must be associated with a pattern of mutually reinforcing organizational structures, business practices, and employee skills that work together as a coherent system. Also, organizational structures have become flatter, decision making has become decentralized, information is widely shared, workers form project teams within and across organizations, and work arrangements are flexible. These changes in organizational structures and practices have been enabled by the application of ICT for communication, information sharing, and simulation of business processes. For example, a U.S. Census Bureau study (Black and Lynch 2003) found significant firm-level productivity increases associated with changes in business practices that included reengineering, regular employee meetings, the use of self-managed teams, up-skilling of employees, and the use of computers by front-line workers. In Canada, Zohgi et al. (2007) found a strong positive relationship between both information sharing and decentralized decision making and a company's innovativeness. Recent studies of firms (Pilat 2004; Gera and Gu 2004) found significant productivity gains when ICT investments were accompanied by other organizational changes, such as new strategies, new business processes and practices, and new organizational structures. Murphy (2002) found productivity gains when the use of ICT was accompanied by changes in production processes (quality management, lean production, business reengineering), management approaches (teamwork, training, flexible work, and compensation), and external relations (outsourcing, customer relations, networking).

These changes in organizational structure and business practices have resulted in corresponding changes in the hiring practices of companies and the skills needed by workers. A study of labor tasks in workplaces found that, commencing in the 1970s, routine cognitive and manual tasks in the U.S. economy declined and nonroutine analytic and interactive tasks grew (Autor et al. 2003). This finding was particularly pronounced for rapidly computerizing industries. The study found that, as ICT is taken up by a firm, computers *substitute* for workers who perform routine physical and cognitive tasks but they *complement* workers who perform nonroutine problem-solving tasks. Similar results were found in the U.K. and the Netherlands (Borghans and ter Weel 2001; Dickerson and Green 2004), France (Maurin and Thesmar 2004) and Canada (Gera and Gu 2004).

Because repetitive, predictable tasks are readily automated, computerization of the workplace has raised the demand for problem-solving and communications tasks, such as responding to discrepancies, improving production processes, and coordinating and managing the activities of others. In a survey of U.K. firms, Dickerson and Green (2004) found an increased demand for technical know-how and for skills in high-level communication, planning, client communication, horizontal communication, problem solving, and checking. Meanwhile, there was a decreased demand for physical skills. The net effect of these changes is that companies in the USA, the UK, and other advanced economies (Lisbon Council 2007) are hiring workers with a higher skill set. It is also interesting that many of these skills (e.g., communication, collaboration, flexibility) are often referred to as "soft skills," yet are some of the most important for success and some of the most difficult to help people develop to high levels of refinement.

The creation of knowledge as a social product (Scardamalia and Bereiter 2003, 2006) is a major part of that higher skill set. It requires collective responsibility for accomplishments, and it is something that scientists, scholars, and employees of highly innovative companies do for a living (Nonaka and Takeuchi 1995). An interesting example is the design of Boeing 787 aircraft, built by nearly 5,000 engineers (not counting production workers) from around the world. The design and engineering work takes place simultaneously at multiple sites, over a long period of time, and yet all the parts ultimately fit nicely together (Gates 2005). In collaborative, creative endeavors of this nature, team members need to understand the top-level goal and share responsibility for the interrelated network of ideas, subgoals, and designs, with success dependent on all members rather than concentrated in the leader. They share responsibility for establishing effective procedures, for assigning and completing practical tasks, for understanding and facilitating team dynamics (Gloor 2006), for remaining cognitively on top of activities and ideas as they unfold (Leonard-Barton 1995), and for the process as a whole. As issues emerge, they collectively shape the next steps, build on each other's strengths, and improve their ideas and designs. Members create the cultural capital of their organization as they refine the "knowledge space" and products that represent their collective work.

Of course this work includes timelines, specified goals, and deadlines. The idea of collective responsibility is not to ignore such aspects but to engage participants in setting deadlines, taking responsibility for achieving them, and redefining goals and schedules as necessary. It also requires a commitment to working in public spaces, making one's thinking and processes explicit and available, and entering artifacts into the shared knowledge space to advance the state of knowledge of the community. If everyone is doing the same thing (as is often the case in schools), the redundant, repetitive work interferes with productivity. The shared problem space needs to grow, based on shared goals and helpful, diverse contributions from all members.

This cluster of changes—organizational structure, business practices, and more-complex employee tasks and skills—is particularly pronounced for knowledge-intensive, knowledge-creating organizations. Probably the most intensive knowledge-creating organizations are research laboratories. Current research in the sociology and anthropology of science has focused on two aspects of the work of scientists: the distributed nature of scientific work over time, resources, and place and the moment-by-moment coordination of instruments, representations, and discourse as scientists construct meaning from the results of their research.

In contemporary science, creating new knowledge requires the coordination of activities through time and across space to assemble methods, tools, and theories, building on previous findings to conduct new research and generate new knowledge (Fujimura 1992). To achieve this spatial and temporal coordination, scientists develop technological and social systems that support the movement of specialized scientific objects, like ideas, data, sketches, and diagrams, across this distributed network. This coordination within and across organizations and across time, place, and objects was apparent in Kozma's study (Kozma et al. 2000; Kozma 2003) of chemists in a pharmaceutical company. Here the synthetic products of one group were frequently the starting materials of another group, as activities related to the creation of a new drug were distributed across laboratories, chemists with different specializations, and equipment with different purposes. This coordination was maintained, in part by standardized procedures and in part by attaching labels with diagrams of chemical structures to the vials as they moved from lab to lab.

The laboratory is where the moment-by-moment work of science is done, much of it centered on instruments and representations. In their collaborative activities, scientists talk and represent visually their ideas to one another in supportive physical spaces (Ochs et al. 1996). The indexical properties of these physical spaces and representations are essential for the ways that scientists collaborate and establish shared meaning (Goodwin and Goodwin 1996; Hall and Stevens 1995; Suchman and Trigg 1993). In their discourse, scientists make references to the specific features of diagrams and data visualizations as they coordinate these representations to understand the products of their work (Kozma et al. 2000; Kozma 2003). The features of these representations are often used as warrants for competing claims about their finding, as scientists try to adjudicate their different interpretations.

These research findings on the practices, organizational structures, and needs of innovative, knowledge-creating organizations have significant implications for the practices and organizational structures of environments needed to support the acquisition of twenty-first century skills and for finding productive connections between in- and out-of-school learning environments. Knowledge-creating organizations rank high on all of the twenty-first century skills listed in various documents and articles (for example, The Partnership for 21st century skills 2009; Binkley et al. 2009; Johnson 2009). Consequently, an analysis of knowledge-creating organizations additionally provides high-end benchmarks and models to guide the design and implementation of modern assessment. For example, the literature on how distributed teams have managed to successfully produce more and better outputs helps to operationalize concepts such as collaboration, group problem solving, use of ICT, and so on. Also relevant are the social, material, and technological practices and organizational structures in which members of knowledge-creating organizations operate.

Table 5.1 maps in condensed form the characteristics of knowledge-creating organizations onto the twenty-first century skills presented in Chap. 1. Our goal is to align these different perspectives and, as elaborated below, provide an analytic framework for educational environments and assessments to identify those most in keeping with characteristics of knowledge-creating organizations.

There are major differences between twenty-first century skills as they figure in school curricula and the skills manifested in knowledge-creating organizations.

Table 5.1 Twenty-first century skills as experienced in knowledge-creating organizations

Twenty-first century skills	Experience in knowledge-creating organizations
Creativity and innovation	Work on unsolved problems; generate theories and models, take risks, etc.; pursue promising ideas and plans
Communication	Knowledge building/progressive discourse aimed at advancing the state of the field; discourse to achieve a more inclusive, higher-order analysis; open community knowledge spaces encourage peer-to-peer and extended interactions
Collaboration/teamwork	Collective or shared intelligence emerges from collaboration and competition of many individuals and aims to enhance the social pool of existing knowledge. Team members aim to achieve a focus and threshold for productive interaction and work with networked ICT. Advances in community knowledge are prized, over-and-above individual success, while enabling each participant to contribute to that success
Information literacy/ research	Going beyond given information; constructive use of and contribution to knowledge resources to identify and expand the social pool of improvable ideas, with research integral to efforts to advance knowledge resources and information
Critical thinking, problem solving, and decision making	High-level thinking skills exercised in the course of authentic knowledge work; the bar for accomplishments is continually raised through self-initiated problem finding and attunement to promising ideas; participants are engaged in complex problems and systems thinking
Citizenship—local and global	Citizens feel part of a knowledge-creating civilization and aim to contribute to a global enterprise; team members value diverse perspectives, build shared, interconnected knowledge spanning formal and informal settings, exercise leadership, and support inclusive rights
ICT literacy	ICT integrated into the daily workings of the organization; shared community spaces built and continually improved by participants, with connection to organizations and resources worldwide
Life and career skills	Engagement in continuous, "lifelong," and "life-wide" learning opportunities; self-identification as a knowledge creator, regardless of life circumstance or context
Learning to learn/ metacognition	Students and workers are able to take charge at the highest, executive levels; assessment is integral to the operation of the organization, requiring social as well as individual metacognition
Personal and social responsibility—incl. cultural competence	Team members build on and improve the knowledge assets of the community as a whole, with appreciation of cultural dynamics that will allow the ideas to be used and improved to serve and benefit a multicultural, multilingual, changing society

In schools the skills are frequently treated separately, each having its own learning progression, curriculum, and assessment. In knowledge-creating organizations different facets of work related to these skills represent a complex system, with the skills so intertwined that any effort to separate them in contexts of use would undercut the dynamic that gives them meaning.

Characteristics of Knowledge-Building Environments

Knowledge-building environments represent *complex systems* that support *emergent outcomes*. They are places that, like knowledge-creating organizations, produce public knowledge—knowledge that does not just reside in the minds of individuals but that is available to others to build on and improve. Public knowledge develops through discourse, in which declarative statements play a necessary role, as do models, theories, and artifacts that are available to the community as a whole. Having students become active agents in knowledge construction is an important theme in the literature on school reform and knowledge-building processes (Engle and Conant 2002; Herrenkohl and Guerra 1998; Lamon et al. 1996; Lehrer et al. 2000; Paavola and Hakkarainen 2005; Tabak and Baumgartner 2004). Of particular interest in this regard is *collective cognitive responsibility*, the requirement to take responsibility for the state of public knowledge (Scardamalia 2002).

As the Boeing example suggests, networked, communal knowledge spaces are at the heart of work in knowledge-creating organizations. Accordingly, the work of participants has an "out-in-the-world" existence. The intellectual life of the community—objectified as theories, inventions, models, plans, and the like—is accessible, in tangible form. In the business world, this is referred to as the organization's corporate knowledge; in the knowledge-building literature, it is referred to as "community knowledge" (Scardamalia 2002). This community knowledge space is typically absent from classrooms, making it hard for students' ideas to be objectified, shared, examined, improved, synthesized, and used as "thinking devices" (Wertsch 1998) so as to enable further advances. It also makes assessment difficult because students' ideas are neither explicit nor in tangible form. In contrast, the commitment to work in open, shared spaces not only renders ideas as objects of discussion and improvement but opens the door for concurrent, embedded, and transformative assessment, as we elaborate below. In turn, these communities can sustain work at the high end of twenty-first century skills, as identified in Table 5.1.

Group Learning

Group learning and group cognition may well become the dominant themes of technology in the next quarter-century, just as collaborative learning was in the previous one (Stahl 2006). Group learning is learning *by* groups, which is not the same as learning *in* groups or individual learning through social processes. The term *learning organization* (Senge 1990) reflects this emphasis on the organization itself operating as a knowledge-advancing entity and reflects the larger societal interest in knowledge creation. Knowledge building is a group phenomenon, even when contributions come from identifiable individuals. Members are responsible for the production of public knowledge that is of value to a community. Again, this maps directly onto the Boeing example presented above. The community may be a research or design group or the world at large, or it may be a group of learners—in which case it is important to distinguish individual learning from the group's knowledge-building accomplishments. Neither one can be reliably inferred from the other, although the interaction between the two is vital and deserving of study in its own right. We return to this issue in the final sections of this chapter.

In a knowledge-building group, the crucial assessment questions are about the group's achievements in advancing the state of knowledge—comparable to the "state of the art" reviews common in the disciplines and professions. Self-assessment by a knowledge-building group can be valuable both for helping the group progress and for individual learning (Lee et al. 2006). External assessment can serve the purposes of troubleshooting and management. Evidence available suggests that such an approach increases individual learning, not just group learning, because the group needs each individual's contribution; thus there is social pressure to perform (e.g., Barron 2003). However, this is a finding much in need of replication and extended study.

Knowledge-Building Developmental Trajectory

Building on the characteristics of knowledge-creating organizations and what we know about learning, we can begin to specify the characteristics of knowledge-building environments and the implications they have for educational practices. Table 5.2 is an elaboration of Table 5.1 and provides a developmental framework for

Table 5.2 Developmental trajectory for knowledge-creating environments

	Characteristics of knowledge-creating organizations	
Twenty-first century skills	Entry level	High
Creativity and innovation	Internalize given information; beliefs/actions based on the assumption that someone else has the answer or knows the truth	Work on unsolved problems; generate theories and models, take risks, etc.; pursue promising ideas and plans
Communication	Social chitchat; discourse that aims to get everyone to some predetermined point; limited context for peer-to-peer or extended interactions	Discourse aimed at advancing the state of the field and at achieving a more inclusive, higher-order analysis; open spaces encourage peer-to-peer and extended interactions
Collaboration/teamwork	Small group work: divided responsibility to create a finished product; the whole is the sum of its parts, not greater than that sum	Shared intelligence from collaboration and competition enhances existing knowledge. Individuals interact productively and work with networked ICT. Advances in community knowledge are prized over individual success, while enabling each to contribute to it

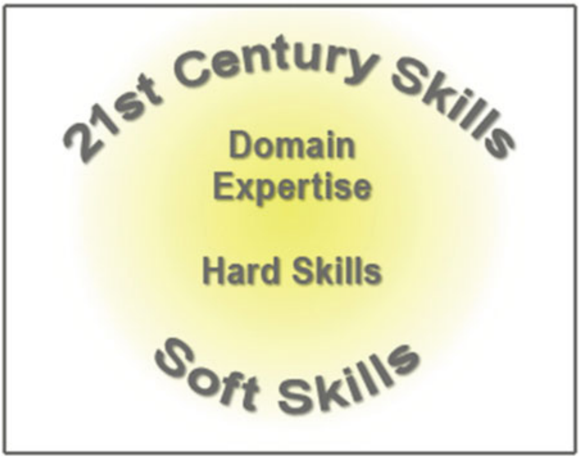
(continued)

Table 5.2 (continued)

	Characteristics of knowledge-creating organizations	
Twenty-first century skills	Entry level	High
Information literacy/research	Inquiry: question-answer, through finding and compiling information; variable testing research	Collaborative expansion of social pool of improvable ideas, with research integral to efforts to advance knowledge
Critical thinking, problem solving, and decision making	Meaningful activities are designed by the director, teacher, or curriculum designer; learners work on predetermined tasks set by others	High-level thinking skills exercised in authentic knowledge work; the bar for accomplishments is continually raised by participants as they engage in complex problems and systems thinking
Citizenship—local and global	Support of organization and community behavioral norms; "doing one's best"; personal rights	Citizens feel part of a knowledge-creating civilization and aim to contribute to a global enterprise; they value diverse perspectives, build shared knowledge in formal and informal settings, exercise leadership, and support inclusive rights
ICT literacy	Familiarity with and ability to use common applications and web resources and facilities	ICT integrated into organization's daily work; shared community spaces built and continually improved by participants, with connection worldwide
Life and career skills	Personal career goals consistent with individual characteristics; realistic assessment of requirements and probabilities of achieving career goals	Engagement in continuous, "life-long," and "life-wide" learning opportunities; self-identification as a knowledge creator, regardless of life circumstance or context
Learning to learn/metacognition	Students and workers provide input to the organization, but the high-level processes are under the control of someone else	Students and workers are able to take charge at the highest, executive levels; assessment is integral to the operation of the organization, requiring social as well as individual metacognition
Personal and social responsibility—incl. cultural competence	Individual responsibility; local context	Team members build on and improve the knowledge assets of the community, with appreciating cultural dynamics that allow the ideas to be used and improved for benefit of multicultural, multilingual, changing society

analyzing learning environments. For each twenty-first century skill, the table suggests a continuum running from the entry-level characteristics that may be expected of students who have had no prior engagement in knowledge building to a level characteristic of productive participants in a knowledge-creating enterprise.

Fig. 5.1 Centrality of deep disciplinary knowledge to all knowledge work



The continuum is an "emergence" continuum—a developmental trajectory from active or constructivist learning as the entry point, to complex systems of interactivity and knowledge work that enable the generation of new knowledge, the capacity to exceed standards, and the drive to go beyond best practice at the high end.

In the section on needed research, we propose experiments to develop this scheme, including additional points along the continuum, to indicate how designing environments with sights set on the high-end of the scale can facilitate the advancement of any school, any teacher along these lines.

Advancing Domain Knowledge and Twenty-First Century Skills in Parallel

Twenty-first century skills—often labeled "soft" or "generic" skills—have been widely recognized as central to innovative capacity and hence as vital for success in a twenty-first century global economy. Although twenty-first century skills are recognized in recent curriculum standards, the main emphasis in standards and assessments is on "hard" skills in language and mathematics as well as "hard" factual knowledge. There is a concern that attention given to "soft" skills will detract from efforts to improve the skills and subject-matter knowledge for which the schools are held accountable. The consensus among researchers in the learning sciences is that these two are not in conflict (Bransford et al. 2000; Darling-Hammond et al. 2008); their interdependence is suggested in Fig. 5.1. In formal education beyond the most basic "3 Rs" level, hard skills are generally treated as a part of domain knowledge. Ability to solve quadratic equations, for instance, is part of algebraic domain knowledge. Hence, as modeled in Fig. 5.1, domain knowledge and hard skills are combined to constitute the focus of formal education, while a common set of soft skills surrounds expertise in all domains.

Making twenty-first century skills universally accessible, rather than the province of knowledge élites, requires that the environments that support knowledge creation be made accessible to all. From the *emergence* perspective, the challenge is to shift to environments that take advantage of what comes naturally to students across the full range of twenty-first century skills (idea production, questioning, communication, problem solving, and so forth) and engage them in the kinds of environments for sustained idea development that are now the province of knowledge élites. These knowledge-building environments that score at the high end of all the developmental continua identified in Table 5.2 increase innovative capacity through engagement in a knowledge-building process—the production of public knowledge of value to others so that processes of collective responsibility for knowledge advancement can take hold (Scardamalia and Bereiter 2003). That is how idea improvement, leading to deep disciplinary knowledge, gets to the center of the enterprise, with twenty-first century skills inseparable and serving as enablers.

Comparative research and design experimentation are needed to add substantially to the knowledge base on relations between inquiry and knowledge-building activities and the meeting of traditional achievement objectives. The research and design experiments proposed in the final section should help address these issues through use of formative assessment, combined with other assessments, selected to evaluate advances in both "hard" and "soft" skills, and the changes over time that are supported through work in information-rich, knowledge-building environments. The proposition to be tested is: *Collective responsibility for idea improvement in environments that engage all students in knowledge advancement should result in advances in domain knowledge in parallel with advances in twenty-first century skills*. This argument is in line with that set forth by Willingham (2008): "Deep understanding requires knowing the facts AND knowing how they fit together, seeing the whole."

This notion that deep understanding or domain expertise and twenty-first century skills are inextricably related has led many to argue that there is not much new in twenty-first century skills—deep understanding has always required domain understanding and collaboration, information literacy, research, innovation, metacognition, and so forth. In other words, twenty-first century skills have been "components of human progress throughout history, from the development of early tools, to agricultural advancements, to the invention of vaccines, to land and sea exploration" (Rotherham and Willingham 2009).

But is it then also true that there are no new skills and abilities required to address the needs of today's knowledge economy? One defensible answer is that the skills are not new but that their place among educational priorities is new. According to Rotherham and Willingham, "What's actually new is the extent to which changes in our economy and the world mean that collective and individual success depends on having such skills. … If we are to have a more equitable and effective public education system, skills that have been the province of the few must become universal." "What's new today is the degree to which economic competitiveness and educational equity mean these skills can no longer be the province of the few" (Rotherham 2008). Bereiter and Scardamalia (2006) have argued, however, that "there is in fact one previously unrecognized ability requirement that lies at the very heart of the knowledge economy. It is the ability to work creatively with knowledge per se." Creative work with knowledge—with conceptual artifacts (Bereiter 2002)—must advance along with work with material artifacts. Knowledge work binds hard and soft skills together.

The deep interconnectedness of hard and soft skills has important implications for assessment, as does the commitment to individual contributions to collective works. As Csapó et al. state in Chapter 4 of this book, "how a domain is practiced, taught, and learned impacts how it should be assessed… the real promise of technology in education lies in its potential to facilitate fundamental, qualitative changes in the nature of teaching and learning" (Panel on Educational Technology of the President's Committee of Advisors on Science and Technology 1997, p.33). Domains in which it is most important to include technology in the assessment of twenty-first century skills include, according to Csapó and colleagues, those in which technology is so central to the definition of the skill that removing it would render the definition meaningless (e.g., the domain of computer programming), those in which higher levels of performance depend on technology tools, and those that support collaboration, knowledge building, and the social interactions critical for knowledge creation. We would argue that to make knowledge building and knowledge creation broadly accessible, technological support for knowledge building also needs to be broadly accessible (e.g., see also Svihla et al. (2009)).

Assessment of "soft" skills is inherently more difficult than assessing the "hard" skills that figure prominently in educational standards. Assessing knowledge-creation processes may be even harder. Nonetheless, this core capability should be further enhanced and clarified through programs of research and design that aim to demonstrate that the processes that underlie knowledge creation also underlie deep understanding; knowledge-building environments promote both. We return to these ideas below.

Advancing Literacy and Closing Gaps

Among the skills needed for life in the knowledge age, literacy is perhaps the most crucial. Without the ability to extract and contribute useful information from complex texts, graphics, and other knowledge representations, one is in effect barred from knowledge work. Print literacy (as with other literacies) has both hard-skill and soft-skill components; e.g., in reading, fluent word recognition is a testable hard skill, whereas reading comprehension and critical reading are important soft skills. Soft-skill components of reading are mandated and tested, but traditional schooling typically deals with them through often ineffectual "practice makes perfect" approaches.

Although there are diverse approaches to literacy education, most of them treat it as an objective to be pursued through learning activities that have literacy as their main purpose. For the most part, with school-based reading, motivation comes from the level of interest in the reading material itself. Consequently, the unmotivated reader, who is frequently one for whom the decoding of print is not fluent, is a persistent problem (Gaskin 2005). During the past decade, however, new approaches have developed in which the focus is not on literacy as such but on collaborative inquiry, where the primary motivation for reading is solving shared problems of understanding. Effects on literacy have been as great as or greater than those of programs that emphasize literacy for its own sake (Brown and Campione 1996; Sun et al. 2008, 2010). Work in Knowledge Forum technology, specially developed to support knowledge building, has provided evidence of significant literacy gains through ICT (Scardamalia et al. 1992; Sun et al. 2008, 2010). Whereas literacy-focused programs typically engage students with reading material at or below their grade level, students pursuing self- and group-directed inquiry frequently seek out material that is above their grade level in difficulty, thus stretching their comprehension skills and vocabularies beyond those normally developed. Rather than treating literacy as a prerequisite for knowledge work, it becomes possible to treat knowledge work as the preferred medium for developing the literacies that support it, with student engagement involving a full range of media objects, so as to support multiliteracies. This approach raises major research issues, which we return to in the final section of this chapter.

Knowledge-Building Analytic Framework

We have developed a *knowledge-building analytic framework* to advance the two goals presented in the introduction to this chapter, to:

- Derive an analytic framework for analyzing environments and assessments that characterize and support knowledge-creating organizations and the knowledge-building environments that sustain them
- Apply this framework to a set of environments and assessments to better understand models, possibilities, and variations in the extent for which they engage students in knowledge-creating organizations or prepare them for work in them

In the "Annex" at the end of this chapter we have included a template that can serve as a scoring scheme to apply to a broad range of environments and assessments, making it possible to characterize strengths and weaknesses of knowledge-building environments and assessments. The scheme is the same as presented above, in Table 5.1. It is simply set up in the "Annex" as a scoring scheme to encourage users to assess specific environments and compare scores by different assessors of the same environment. Users have reported that it is a helpful instrument for reflection on key aspects of the environment analyzed, and becomes increasingly beneficial once they have a chance to view and discuss ratings of the same environment by different raters. The discussion of rationales for different ratings facilitates understanding of the dimensions and functions associated with knowledge-creating organizations. Graduate students studying in the field of knowledge creation tended to rate environments lower than the proponents of those environments (see Table 5.3 and Fig. 5.6, second section of the "Annex"), but not much can be made of this, as the sample is very small. We offer the template to foster the sort of conversation that may be engendered through analysis of a developmental framework related to characteristics of a knowledge-creating organization.

Knowledge-Building and Learning Theories

An important question is how competencies that foster work in a knowledge society relate to modern theories of learning. For example, how does an emphasis on knowledge building fit the "How People Learn" framework, shown in Fig. 5.2, which has been used by a National Academy of Science committee to organize what is known about learning and teaching (National Research Council 2000). The framework highlights a set of four lenses that can be used to analyze learning environments, ranging across homes, community centers, classrooms, schools, and higher levels of

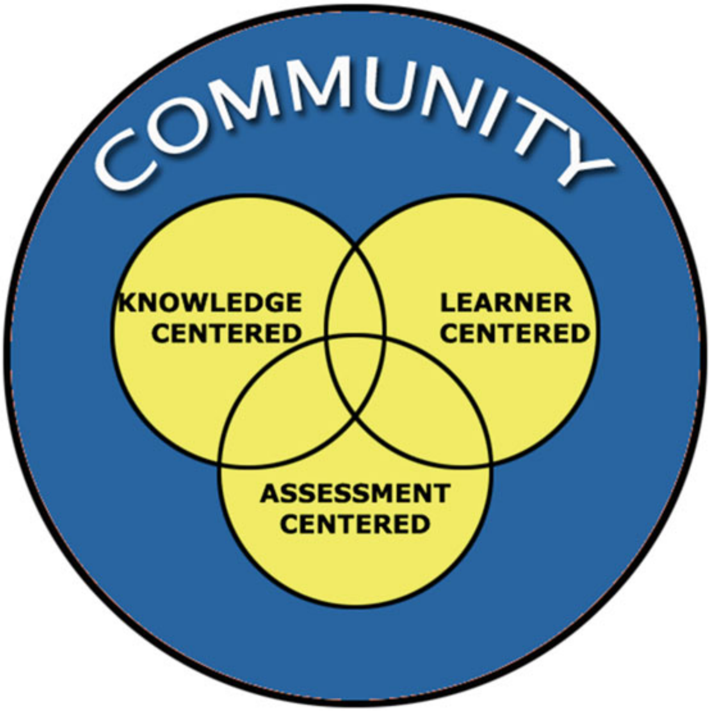


Fig. 5.2 The "How People Learn" framework (Adapted from How People Learn–National Research Council, 2000)

educational organization. The components of the framework involve a focus on four areas that need to be flexibly balanced, depending on current goals and needs. Each area of the framework is accompanied by a set of questions that are useful for exploring the design of learning opportunities, particularly those that support knowledge building.

1. *Knowledge centered*: What needs to be taught to meet the changing needs of people and societies? (Answering this question is fundamental to this entire project.)
2. *Learner centered*: How can new information be connected with learners' existing beliefs, values, interests, skills, and knowledge so that they learn with understanding and can flexibly use what they know?
3. *Community centered*: How can we develop communities of learners that value excellence as people work together to build new knowledge for the common good? And how can we broaden our sense of community and explore opportunities for learning that connect activities in and outside schools?
4. *Assessment centered*: How can we develop frequent and useful opportunities for students, teachers, school systems, and nations to assess the progress they are making toward twenty-first century skills?

Knowledge Centered

As discussed above, the world has changed and different kinds of skills and knowledge are required for successful and productive lives in the twenty-first century. Many of the skills identified above are not tied directly to traditional subject domains, such as the sciences, mathematics, or history—all these, of course, will continue to be important in the twenty-first century. Work by contributors to this series of chapters suggests that constant questioning about what people need to learn is one of the most important activities for our future.

Expertise and Knowledge Organization

More than ever before, experts' knowledge must be more than a list of disconnected facts and must be organized around the important ideas of current and expanding disciplines. This organization of knowledge must help experts know when, why, and how aspects of their vast repertoire of knowledge and skills are relevant to any particular situation (see Bransford et al. 2000). Knowledge organization especially affects the ways that information is retrieved and used. For example, we know that experts notice features of problems and situations that may escape the attention of novices (e.g., see Chase and Simon 1973; Chi et al. 1981; de Groot 1965). They therefore "start problem solving at a higher place" than novices (de Groot 1965). Knowledge building suggests that learning must include the desire and ability to notice new connections and anomalies and to actively seek ways to resolve disconnects by restructuring what they know and generating new, domain-bridging ideas.

Generative knowledge building must also be structured to transcend the problem that current courses and curriculum guidelines are often organized in ways that fail to develop the kinds of connected knowledge structures that support activities such as effective reasoning and problem solving. For example, texts that present lists of topics and facts in a manner that has been described as "a mile wide and an inch deep" (e.g., see Bransford et al. 2000) are very different from those that focus on the "enduring ideas of a discipline" (Wiske 1998; Wilson 1999). However, a focus on knowledge building goes beyond attempts to simply improve learning materials and seeks to help learners develop the vision and habits of mind to develop their own abilities to refine, synthesize, and integrate.

Adaptive Expertise

An especially important focus on knowledge building separates "routine experts" from "adaptive experts" (e.g., Hatano and Inagaki 1986; Hatano and Osuro 2003). Both routine experts and adaptive experts continue to learn throughout their lifetimes. Routine experts develop a core set of skills that they apply throughout their lives with greater and greater efficiency. In contrast, adaptive experts are much more likely to change their core skills and continually expand the breadth and depth of their expertise. This restructuring of core ideas, beliefs, and skills may reduce their efficiency in the short run but make them more flexible in the long run. These processes of restructuring often have emotional consequences that accompany realizations that cherished beliefs and practices need to be changed. Research by Anders Ericsson and colleagues (2009) shows that a major factor in developing expertise is to resist plateaus—in part by continually moving out of one's comfort and engaging in "deliberate practice." This analysis of expertise highlights the need for unlearning as well as learning, and for the kinds of social collaboration that are often invisible when we see write-ups of "experts" in the research literature or the media (e.g., see Bransford and Schwartz 1999).

This research has implications for the design of environments to support knowledge building. First, an emphasis on building a deep understanding of key ideas is important. This serves as the basis for organizing facts that would otherwise depend on sheer memorization. Second, understanding with respect to the adaptability of knowledge structures highlights the need to support processes of review and reflection.

Learner Centered

The learner-centered lens of the How People Learn framework overlaps with the knowledge-centered lens, but specifically reminds us to think about learners rather than only about subject matter. Many educators deal with issues of understanding learners in ways that allow them to engage in culturally responsive teaching (e.g., Banks et al. 2007). This includes learning to build on people's strengths rather than simply seeing weaknesses (e.g., Moll 1986a, b), and helping people learn to "find their strengths" when confronted with new knowledge building challenges. Several important aspects of being learner centered are discussed below.

Understanding the Constructive Nature of Knowing

The constructive nature of knowing grew out of the work of Swiss psychologist Jean Piaget. Piaget used two key terms to characterize this constructive nature: *assimilation* and *accommodation*. In Piaget's terms, learners assimilate when they incorporate new knowledge into existing knowledge structures. In contrast, they accommodate if they change a core belief or concept when confronted with evidence that prompts such as change.

Studies by Vosniadou and Brewer illustrate assimilation in the context of young children's thinking about the earth. They worked with children who believed that the earth is flat (because this fit their experiences) and attempted to help them understand that, in fact, it is spherical. When told it is round, children often pictured the earth as a pancake rather than as a sphere (Vosniadou and Brewer 1989). If they were then told that it is round like a sphere, they interpreted the new information about a spherical earth within their flat-earth view by picturing a pancake-like flat surface inside or on top of a sphere, with humans standing on top of the pancake. The model of the earth that they had developed—and that helped them explain how they could stand or walk upon its surface—did not fit the model of a spherical earth. Everything the children heard was incorporated into their preexisting views.

The problem of assimilation is relevant not only for young children but also for learners of all ages. For example, college students have often developed beliefs about physical and biological phenomena that fit their experiences but do not fit scientific accounts of these phenomena. These preconceptions must be addressed in order for them to change their beliefs (e.g., Confrey 1990; Mestre 1994; Minstrell 1989; Redish 1996). Creating situations that support accommodation is a significant challenge for teachers and designers of learning environments—especially when knowledge building is involved.

Connecting to Students' Previous Experiences

Ideally, what is taught in school builds upon and connects with students' previous experiences, but this is not always the case. A number of researchers have explored the benefits of increasing the learner centeredness of teaching by actively searching for "funds of knowledge" in students' homes and communities that can act as bridges for helping them learn in school (e.g., Lee 1992; Moll 1986a, b; Moses 1994). Examples include helping students see how the carpentry skills of their parents relate to geometry, how activities like riding the subway can provide a context for understanding algebra, and how everyday language patterns used outside of school often represent highly sophisticated forms of language use that may be taught in literature classes as an academic subject but have not been linked to students' out-of-school activities. Work by Bell and colleagues specifically links activities in homes and communities with work in schools (e.g., Bell et al. 2009; Tzou and Bell 2010).

Learner Centeredness, Metacognition, and Basic Cognitive Processes

Being learner centered also involves an awareness of some basic cognitive processes that influence learning for everybody. "Metacognition" is the field of psychology that can be used to help people learn about the cognitive processes that underlie their own abilities to learn and solve problems. Several cognitive processes are particularly important.

Attention and Fluency

Learning about attention is an important part of becoming a metacognitive learner. For example, there are important constraints on how much we can pay attention to at any particular point in time. The amount of attention that we need to devote to a task depends on how experienced and efficient we are at doing it. When learning to read, for example, the effortful allocation of attention to pronouncing words can make it difficult to also attend to the meaning of what one is reading. The attentional demands that accompany attempts to learn anything new mean that all learners must go through a period of "klutziness" as they attempt to acquire new skills and knowledge. Whether people persist or bail out during these "klutz" phases depends in part on their assumptions about their own abilities. Some people may decide "I'm not good at this" and give up trying before they have a chance to learn effectively (e.g., Dweck 1986). Wertime (1979) notes that an important part of being learner centered is to help students learn to persist in the face of difficulty by increasing their "courage spans."

Technology presents challenges of "multitasking," and many students feel that this does not hurt their performance. They can be helped to test this idea for themselves by listening to a lesson with full attention versus listening to one while also multitasking. This is an effective way to help students discover their own abilities and limits rather than simply be forced to comply with "no computers can be on in this class."

Transfer

Learning about ourselves as learners also involves thinking about issues of transfer—of learning in ways that allow us to solve novel problems that we may encounter later.

The mere memorization of information is usually not sufficient to support transfer. Learning with understanding typically enhances the experience (e.g., NRC 2000). An important goal for transfer is cognitive flexibility (e.g., Spiro et al. 1991). Experts possess cognitive flexibility when they can evaluate problems and other types of cases in their fields of expertise from many conceptual points of view, seeing multiple possible interpretations and perspectives. Wiggins and McTighe (1997) argue that understanding complex issues involves being able to explain them in more than one way. Spiro et al. (1991) argue that the inability to construct multiple interpretations in analyzing real-world cases can result from instruction that oversimplifies complicated subject matter.

Motivation

Helping students learn to identify what motivates them is also an important part of being learner centered that contributes strongly to knowledge building. Researchers have explored differences between extrinsic motivators (grades, money, candy, etc.) and intrinsic motivators (wanting to learn something because it is relevant to what truly interests you). Both kinds of motivation can be combined; for example, we can be intrinsically interested in learning about some topics *and* interested in receiving extrinsic rewards as well (e.g., praise for doing well, a consultanting fee). However, some people argue that too much of an emphasis on extrinsic rewards can undermine intrinsic motivation because people get too used to the external rewards and stop working when they are removed (e.g., Robinson and Stern 1997).

There appear to be important differences between factors that are initially motivating (the assumption that learning to skateboard seems interesting), and factors that *sustain* our motivation in the face of difficulty ("hmm, this skateboarding is harder to learn than it looked"). The social motivation support of peers, parents, and others is an especially important feature that helps people persist in the face of difficulties. It is also important to be provided with challenges that are just the right level of difficulty—not so easy that they are boring and not so difficult that they are frustrating. Creating the right kinds of "just manageable difficulties" for each student in a classroom constitutes one of the major challenges and requires expert juggling acts. Explorations of the literature on motivation can be found in Deci and Ryan (1985), Dweck (1986) and Stipek (2002).

Agency

An emphasis on knowledge building especially highlights an important aspect of metacognition and motivation that involves the need for people to develop socially responsive agency. That is, students must learn to make their own choices, experience the social consequences that arise from them, and revise their strategies when necessary. This is a progressive process of moving from the situation in which the teacher makes decisions about student learning to one where students are increasingly responsible for their own learning activities.

An example involves a recent set of studies on science kits for middle school students (Shutt et al. 2009). They involve hands-on activities such as working with and studying (without harming them) fish, isopods, and a variety of other creatures. Throughout the course of the year, the goal is to develop a sense of key variables (e.g., range of temperatures, ranges of acidity, etc.) that affect the life of all species. As originally developed, the science work is extremely teacher directed; the hypotheses to be tested and the methods to be used, such as determining whether isopods desire moist or dry soil, are specified by the teacher. Redesigning these teaching situations has been found to give much more agency to the students. They are given a terrarium and told that their task (working in groups) is to keep their organisms (e.g., isopods) alive. To be successful, they have to choose what questions to ask, how to run the studies, how to do the kind of background research (via technology when needed), and so forth. The initial findings (more precise data will be available soon) show that the sense of agency is very important to students and they take their work very seriously. This kind of activity can hopefully strengthen other skills such as global sensitivity since the students all do their work with the well-being of others (even though they are nonhumans) foremost in their minds.

Community Centered

The preceding discussion explored a number of issues relevant to being knowledge centered and learner centered. The community centered aspect of the How People Learn framework is also related to being knowledge and learner centered, but it focuses special attention on the social, material, and temporal nature of learning.

The Social Aspects of Learning

The social aspects of learning often include the norms and modes of operation of any community that we belong to or are joining. For example, some classrooms represent communities where it is safe to ask questions and say, “I don’t understand this, can you explain it in a different way?” Others follow the norm of, “Don’t get caught not knowing something.” A number of studies suggest that—in order to be successful—learning communities should provide people with a feeling that members matter to each other and to the group, and a shared belief that members’ needs will be met through their commitment to be together (Alexopoulou and Driver 1996; Bateman et al. 1998). Many schools are very impersonal places, and this can affect the degree to which people feel part of, or alienated from, important communities of professionals and peers.

Concerns that many schools are impersonal and need to be smaller in order to be more learner and community centered can also be misinterpreted as simply being an argument for helping students feel good about themselves. This is very important, of course, but more is involved as well. More includes searching for “funds of knowledge" in students' lives and communities that can be built upon to enhance their motivation and learning. The more we know about people, the better we can communicate with them and hence help them (and us) learn. And the more they know about one another, the better they can communicate as a community.

The importance of creating and sustaining learning communities can be traced to Vygotsky's theory in which culture and human interaction represent central developmental processes. Vygotsky focused on the intersection between individuals and society through his concept of the zone of proximal development (ZPD)—the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance, or in collaboration with more capable peers (Vygotsky, 1962/1934). What a child can perform today with assistance, she will be able to perform tomorrow independently, thus preparing her for entry into a new and more demanding collaboration. The emphasis here is on the ways learners draw on each other for ideas and resources that support or scaffold their own learning.

The Material Aspects of Learning

Vygotsky also emphasized the ways in which material resources, such as tools and technologies, change the nature of tasks and the cognitive skills that are required to perform them. This is particularly important in the twenty-first century, not only because of the ways in which technologies have changed the nature of task and work in the world outside of schools but because students increasingly use a wide range of technologies in their everyday lives and bring these technologies with them into schools. Often teachers do not take advantage of these technologies or use the skills and experiences that students bring with them as a way to increase students' knowledge of school subjects or further develop their twenty-first century skills. Learning and assessment are far different if students have access to a range of technological tools, digital resources, and social support than if they learn or are assessed without access to these resources; while the real world of work and students' social environments are filled with these tools and resources, they can be effectively built into the learning environment (Erstad 2008).

The Temporal Aspects of Learning

At a broader level, being community centered also means reaching beyond the walls of the schools in order to connect with students' out-of-school experiences, including experiences in their homes.

Figure 5.3, from the LIFE Center, illustrates the approximate time spent in formal (school) and informal (out-of-school) environments. A great deal of learning goes on outside of school (Banks et al. 2007), but often teachers do not know how to connect these kinds of experiences to school learning. Earlier we discussed the idea of searching for "funds of knowledge" that exist in communities and can be built

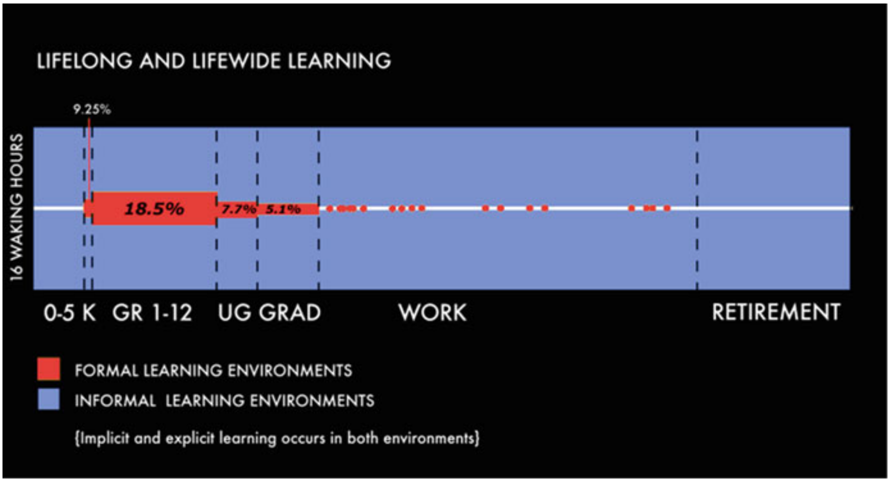


Fig. 5.3 Time spent in formal and informal learning across a typical lifespan. Estimated time spent in school and informal learning environments. Note: This diagram shows the relative percentage of their waking hours that people across the lifespan spend in formal educational environments and other activities. The calculations were made on the best available statistics for a whole year based on how much time people at different points across the lifespan spend in formal instructional environments. (Reproduced with permission of The LIFE Center.) (The LIFE Center's Lifelong and Lifewide Diagram by LIFE Center is licensed under a Creative Commons Attribution-Share Alike 3.0 United States License . (LIFE Center: Stevens et al. 2005) LIFE Center (2005). "The LIFE Center's Lifelong and Lifewide Diagram". This diagram was originally conceived by Reed Stevens and John Bransford to represent the range of learning environments being studied at the Learning in Informal and Formal Environments (LIFE) Center (http://life-slc.org). Graphic design, documentation, and calculations were conducted by Reed Stevens, with key assistance from Anne Stevens (graphic design) and Nathan Parham (calculations)

upon so as to help students succeed. The challenge is to help students build strong social networks within a classroom, within a school, and between classrooms and in- and out-of-school contexts.

Assessment Centered

We've discussed learning centered on knowledge, learner, and community; now we turn to assessment-centered learning. It is easy to assume that assessment simply involves giving tests to students and grading them. Theories of learning suggest roles for assessment that involve much more than simply making up tests and giving grades.

First, teachers need to ask what they are assessing. This requires aligning their assessment criteria with the goals for their students (part of being knowledge centered) and the "readiness" of students in their classroom (learner and community centered). Assessing memorization (e.g., of properties of veins and arteries) is different from assessing whether students are understanding why veins and arteries have various properties. Similarly, assessing whether students can answer questions about life cycles (of frogs, for example) is different from assessing whether they will spontaneously retrieve this information when attempting to solve problems.

At the most general level, issues of what to assess relate to the issue of what students need to know and be able to do in order to have fulfilling lives once they graduate. Because of rapid changes in society, this is an issue that constantly needs to be reconsidered. Debates about standardized tests include concerns that they may "tip" teaching in a direction that is counterproductive for students because some teachers spend most of their time teaching to the tests while the tests do not assess the range of skills, knowledge, and attitudes needed for successful and productive lives in the twenty-first century.

Different Kinds and Purposes of Assessment

An especially important aspect of the assessment-centered lens in the How People Learn framework is its emphasis on different kinds of assessments for different purposes. When most people think about assessments, they think about *summative assessments*. These include unit exams at the end of a unit, standardized tests at the end of the year, and final exams at the end of a course. Summative assessments come in all forms: multiple choice tests, essays, presentations by students, and so forth. These assessments are very important as an accountability mechanism for schools, teachers, and students. Often they reveal important information that the teachers wish they had seen earlier. This is why *formative assessments* are important. These are used for the purpose of improving teaching and learning. They involve making students' thinking visible as they progress through the course, giving them feedback about their thinking, and providing opportunities to revise.

Assessment and Theories of Transfer

It is also important for teachers to understand ways in which assessment practices relate to theories of transfer. Consider summative assessments, for example. We all want to make sure that these provide an indication of students' ability to do something other than simply "take tests." Ideally, our assessments are predictive of students' performance in everyday settings once they leave the classroom.

One way to look at this issue is to view tests as attempts to predict students' abilities to *transfer* from classroom to everyday settings. Different ways of thinking about transfer have important implications for thinking about assessment. Central to traditional approaches to transfer is a "direct application" theory and a dominant methodology that Bransford and Schwartz (1999) call "sequestered problem solving" (SPS). Just as juries are often sequestered in order to protect them from possible exposure to "contaminating" information, subjects in experiments are sequestered during tests of transfer. There are no opportunities for them to demonstrate their abilities to learn to solve new problems by seeking help from other resources, such as texts or colleagues, or by trying things out, receiving feedback, and getting opportunities to revise. Accompanying the SPS paradigm is a theory that characterizes transfer as the ability to directly apply one's previous learning to a new setting or problem. We call this the direct application (DA) theory of transfer. Some argue that the SPS methodology and the accompanying DA theory of transfer are responsible for much of the pessimism about evidence for transfer (Bransford and Schwartz 1999).

An alternative view that acknowledges the validity of these perspectives also broadens the conception of transfer by including an emphasis on people's "preparation for future learning" (PFL). Here, the focus shifts to assessments of people's abilities to learn in knowledge-rich environments. When organizations hire new employees, they don't expect them to have learned everything they need for successful adaptation. They want people who can learn, and they expect them to make use of resources (e.g., texts, computer programs, and colleagues) to facilitate this learning. The better prepared they are for future learning, the greater the transfer (in terms of speed and/or quality of new learning). Examples of ways to "prepare students for future learning" are explored in Schwartz and Bransford (1998), Bransford and Schwartz (1999) and Spiro et al. (1987).

The sole use of static assessments may mask the learning gains of many students, as well as masking the learning advantages that various kinds of educational experiences provide (Bransford and Schwartz 1999). Linking work on summative assessment to theories of transfer may help us overcome the limitations of many existing tests. Examples of SPS versus PFL assessments of learning and transfer are discussed in Bransford and Schwartz (1999).

Implications for Assessment Reform

Two distinct approaches to the design of environments and assessment have been described. One involves working backward from goals to construct a system of subgoals and learning progressions from an initial state to the goal. The second approach involves e*mergent goals* that are not fixed in advance but take shape as learning and thinking proceed. We have indicated the trade-offs associated with both the *working-backward* and *emergence* approaches, and below, after reviewing assessment challenges related to twenty-first century skills, we specify the research needed, depending on what one sets out to pursue. In the additive model the "twenty-first century skills" curriculum is added to the traditional curriculum, although often the goal is more in line with assimilative efforts to merge skill and content elements or to piggyback one upon the other. The problem, exacerbated if each twenty-first century skill is treated separately, is that the current "mile wide, inch deep" curriculum will grow miles wider and shallower, with the twenty-first century skills curriculum taking valuable time away from traditional skills. The goal of the transformational model is to effect a deeper integration of domain understanding with twenty-first century skills. The rationale, elaborated in the section on the parallel advance of domain knowledge and twenty-first century skills, is that if a deep understanding of domain knowledge is achieved through exercising twenty-first century skills, the result will be enhanced understanding in the domain, as well as advances in twenty-first century skills. That is the guiding principle underlying the knowledge-building approach. The knowledge-building analytic framework, described in the "Annex," helps those wishing to engage in this transformation to consider progress along its multiple dimensions. Since these dimensions represent a complex interactive system, treating them separately may prove more frustrating than helpful. Fortunately, this also means that tackling one dimension is likely to lead to advances along several of them. The implication for assessment is that we must anticipate and measure generalization effects. We elaborate possibilities for design experiments to integrate working-backward and emergence models in the section on specific investigations. But first we discuss a broader set of issues regarding assessment challenges and twenty-first century skills.

Assessment Challenges and Twenty-First Century Skills

The quest for evidence-based assessment of twenty-first century skills is hindered by many factors. First, there are huge variations in formal and informal learning environments and the kinds of assessment that are possible in them. Second, the knowledge and skills that deal with the media and technologies used within a domain need to be distinguished from domain-specific knowledge and skills (Bennett et al. 2007; Quellmalz and Kozma 2003). Third, methods for designing twenty-first century assessments and for documenting their technical quality have not been widely used (Quellmalz and Haertel 2008). Fourth, assessments need to be coherent across levels of educational systems (Quellmalz and Pellegrino 2009; Pellegrino et al. 2001). Coherence must start with agreement on the definition of twenty-first century skills and their component knowledge and techniques. Moreover, the design of international-, national-, state-, and classroom-level tests must be clarified and aligned, otherwise assessments at different levels will not be balanced and inferences about student performance will be compromised.

Evidence-centered design (Messick 1994; Mislevy and Haertel 2006) links twenty-first century skills to the task features and reports of evidence that characterize student performance and progress. In the sections immediately following, we describe how evidence-centered design can be used to develop formative assessments that are embedded in learning environments and that link these formative assessments to large-scale, summative assessments.

Cognitively Principled, Evidence-Centered Assessment Design

As described above, research on the development of expertise in many domains has indicated that individuals proficient in a domain have large, organized, interconnected knowledge structures and well-honed domain-specific problem-solving strategies (Bransford et al. 2000). The design of assessments, therefore, should aim to measure both the extent and connectivity of students' growing knowledge structures and problem-solving strategies (Pellegrino et al. 2001; Glaser 1991). For example, in the domain of science, core knowledge structures are represented in models of the world built by scientists (Hestenes et al. 1992; Stewart and Golubitsky 1992). Technologies are seen as tools that support model-based reasoning by automating and augmenting performance on cognitively complex tasks (Norman 1993; Raizen 1997; Raizen et al. 1995).

The NRC report, *Knowing What Students Know,* presents advances in measurement science that support the integration of cognitive research findings into systematic test design frameworks. As a brief overview, evidence-centered assessment design involves relating the learning to be assessed, as specified in a *student model*, to a *task model* that specifies features of the task and questions that would elicit observations of learning, and to an *evidence model* that specifies the student responses and scores that serve as evidence of proficiency (Messick 1994; Mislevy et al. 2003; Pellegrino et al. 2001). These components provide a structure for designing assessments of valued twenty-first century skills and also for evaluating the state of current assessment practices. Evidence-centered design (Messick 1994; Mislevy and Haertel 2006) can be used to design formative assessments and link these to large-scale, summative assessments.

The Role of Domain Knowledge

An issue for large-scale twenty-first century assessments is the role of knowledge about topics and contexts in a discipline or specialization that is required to accomplish tasks and technology-based items. Large-scale assessments of twenty-first century skills cannot assume that all students will have learned a particular academic content. Fortunately, assessments of twenty-first century skills within learning environments *can* identify the content knowledge within which they will be situated. In academic subjects, current assessments of problem-solving and critical-thinking skills, if they are directly assessed and reported at all, are typically reported as components of subject-matter achievement (i.e., math problem solving, science inquiry), not as distinct twenty-first century skills. In addition, in core school subjects as well as informal settings, students may use common or advanced technologies, but their technology proficiencies tend not to be tested or reported. Therefore, to assess and report progress on twenty-first century skills, the design of assessments of students' performance relevant to them must specify the knowledge and skills to be tested and reported for each skill (see Chap. 2); either crosscutting processes such as problem solving or communication, or their ability to use technologies in a range of academic and practical problems. An important feature of knowledge-building environments and the assessments of ICT skills within them will be to test not only the use of ICT tools, simple and advanced, but also the learners' skill in using a range of ICT tools to extend and build their knowledge and strategies for increasingly more complex tasks. In addition, learners' adaptive expertise, their ability to transfer their existing knowledge and strategies to novel problems, will need to include direct assessment of their ability to learn and apply new technologies.

Assessments Embedded in Technology-Rich Environments

The design of assessments must begin by specifying their purposes and intended uses (AERA/APA/NCME 1999). These specifications then lead to validity questions such as "Does the assessment support the inferences and actions based on it?" The two conventional distinctions are between summative and formative purposes. As indicated earlier, summative assessments are administered at the end of an intervention, or a unit within it, so as to judge whether goals have been met. Formative assessments are administered during interventions to inform learners and instructors, giving time for midcourse corrections. A recent definition proposed in the USA by the Formative Assessment for Students and Teachers (FAST) state collaborative, supported by the Council of Chief State School Officers, is that "Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes." According to the FAST definition, formative assessment is not an instrument but the process of using information about progress toward a goal to improve learning. Important attributes of formative assessments are that the outcomes are *intended* and clearly *specified* in advance, the methods are deliberately planned, the *evidence of learning* is used by teachers and students, and *adjustments occur during instruction*. Attributes of effective FAST formative assessment include: clearly articulated learning progressions; learning goals and criteria for success that are clearly identified and communicated to students; evidence-based descriptive feedback; self and peer assessment; and collaboration of students and teachers in working toward learning goals. Formative assessments of twenty-first century skills, therefore, would specify the twenty-first century outcomes and systematic methods for monitoring progress and providing feedback, as well as clear criteria for success. Formative assessments for twenty-first century skills could be employed for all the twenty-first century skills in all kinds of learning environments.

This FAST prescription of formative function of twenty-first century assessments is quite different from the use of embedded assessments to validate large-scale assessment results, or to augment the evidence that could be collected in a one-time, on-demand test. A third function of embedded assessments can be to collect detailed information about processes and progress for research purposes, and to begin to create a more coherent integration of formative and summative assessment.

What Evidence Will Be Sought?

Within an evidence-centered design assessment framework, broad twenty-first century skills, such as problem solving or communication, need to be further dissected into component targets for assessment. Problem-solving targets in mathematics might involve planning solution strategies or evaluating solutions. In science, problem solving might involve targets such as planning investigations or interpreting data in visualizations (Quellmalz and Kozma 2003). In literature, problem solving may involve analyses of Shakespeare plays, looking for recurring symbolism related to the plot. Problem-solving targets to assess in a practical situation might involve selecting a green technology, such a wind turbine, and analyzing its potential environmental impacts. The assessment targets for twenty-first century problem-solving skills will be at a more general level for applications across domains and situations. Problem-solving assessment tasks will need to represent structured problems with known solutions as well as problems with multiple solutions. In domain-centered learning environments, assessment tasks will go beyond the repetition of previously performed experiments to open-ended tasks permitting numbers of appropriate methods for eliciting evidence of how well learners plan, conduct, and interpret evidence in solving a problem or achieving a goal.

Evidence-centered assessment design requires that embedded assessments articulate the qualitative or quantitative information that would document achievement of each twenty-first century skill and its component targets. For formative assessments, a crucial feature is that the evidence and criteria be understandable and useable by teachers and students. For example, self and peer assessment are key features of effective formative assessment. Such activities are already familiar in classes that use peer review of drafts of compositions or peer critiques of presentations. In the workplace, peer review is a hallmark of professional publications.

While common Internet and productivity tools are often integrated across contexts and disciplines, the "tools of the trade" differ between humanities, sciences, and social sciences, and so on, as well as between postsecondary learning environments, the workplace, and the professions. In primary and secondary formal schooling, common Internet and productivity tools are often integrated across contexts and disciplines. Once again, the knowledge and skills will need to be specified and further decomposed as they apply to different learning environments. Evidence of achievement will also need to be specified in ways that are shareable with learners and teachers. Thus embedded assessments of use of specific technologies will vary according to the context and domains emphasized. Nonetheless, new assessment possibilities are opening up through efforts to create tools that are useable across domains and that link domain-specific environments with more general environments.

Twenty-first century skills are difficult to assess with timed, on-demand large-scale tests, and typically better monitored over time within learning environments. For example, creativity and innovation can be assessed in relation to how learners have gone beyond what was specified in learning activities. Collaboration with present and virtual peers and experts can be monitored throughout formation of teams, integration of contributions, and feedback to reflect on the effectiveness of the team processes and the achievement of goals.

Design of Assessments to Elicit Evidence of Twenty-First Century Skills

Systematic, direct assessment of twenty-first century skills in classrooms is rare. Although students may be taught to use common and advanced tools, teachers tend not to have specific standards for twenty-first century skills for students to meet nor testing methods to gather evidence of student skill in using the technologies. In either formal or informal learning environments, teachers are typically left on their own to figure out how to integrate technology into their curricula or into informal learning activities. The state of integration of the assessment of twenty-first century skills into learning activities remains in its infancy.

Assessment must be designed to elicit evidence of learning related to each assessment target. Research on effective formative assessment describes types of formal and informal observations of learning, from questions to and from learners, to examinations of work in progress, and evaluations of work products. However, these observations should be planned for in advance with the criteria for success laid out and shared with learners. For example, systematic observations of groups during collaboration activities can be structured so as to record the types and quality of interactions. These observations can be summarized and reviewed with groups and individuals.

The twenty-first century skills integrate learners' use of a range of technologies over the variety of contexts and domains in the learning environments. Central to the twenty-first century skills is the learner's ability to select and use appropriate technologies during processes such as innovation, communication, collaboration, problem solving, and citizenship. Technologies offer many possibilities for designing richer, deeper, wider-ranging learning activities and assessments. Possibilities for technology-supported reform of learning environments and assessments include:

- Provision of authentic, rich, dynamic environments
- Access to collections of information sources and expertise
- Use of formal and informal forms of collaboration and social networking
- Presentation of phenomena, difficult or impossible to observe and manipulate in classrooms
- Examples of temporal, causal, dynamic relationships "in action"
- Allowing multiple representations of stimuli and their simultaneous interactions (e.g., data generated during a process)
- The use of overlays of representations, symbols
- Student manipulations/investigations, multiple trials
- Student control of pacing, replay, revision
- Making student thinking and reasoning processes visible

- Capturing student responses during activities (e.g., research, design, problem solving)
- Allowing the use of simulations of a range of tools (Internet, productivity, domain based)

Below, in the section on assessment and the knowledge-building developmental trajectory, we extend this list. But first we introduce the notion of an assessment profile and elaborate on the potential for new environments and assessments to inform and be informed by large-scale assessments.

Assessment Profile

The purpose of the knowledge-building analytic framework, (see "Annex") is to determine the extent to which an educational environment is moving toward a knowledge-creating enterprise, in line with the developmental trajectories defined in Table 5.2. The assumption underlying this framework is that educational environments, not only students, should be evaluated. But of course the work of students must also be analyzed, and for this purpose these dimensions need to be translated into measures of individual and group performance. We propose such work as part of a necessary program of research. But for now we offer six dimensions of assessment to support use and coverage of all manner of assessments to measure twenty-first century skills, across all classrooms, so as to ensure quality assessments and to guide instructional practices.

Alignment between assessments and twenty-first century skills. Some assessment instruments may not assess or support one or more of the twenty-first century skills, so it is helpful, for each target twenty-first century skill, to determine if there is (1) full, (2) partial, or (3) no alignment.

Purpose and intended use of assessments. Assessment data, tasks, and items may serve as (1) formative assessments, so students and instructors can monitor learning and adjust instruction as it proceeds; (2) summative evidence of end-of-instruction achievements; or (3) project evaluation or research, not shared with learners and instructors. For each twenty-first century skill, it is worth tracking its purpose on each of these purposes.

Construct representation. Assessment tasks and items can sometimes produce evidence about only portions of the targeted constructs, desired knowledge or skills. For example, if the target is systems knowledge, components or simple interactions may be tested rather than dynamic, emergent behaviors. Or basic facts or steps may be tested rather than higher-level, integrated knowledge and skills. When constructs are only partially tested, important components may not be fully represented. For each twenty-first century skill, it should be determined whether available evidence represents (1) the construct, (2) part of the construct, or (3) none of the construct.

Integration into learning activities. Assessments in learning environments may be integrated into ongoing activities to a greater or lesser extent. Integrated, ongoing assessments can gather evidence of learning throughout their activities. Interim assessments less directly linked to ongoing activities may be periodically administered as checks. Or, decontextualized, external assessments can be dropped in. Thus it is helpful, for each twenty-first century skill, to determine the extent to which tasks and item responses (1) are fully integrated into learning activities; (2) are assessed afterward, separately from learning activities; or (3) are not assessed.

Feasibility. Assessments in learning environments may also differ in the feasibility of their use. They may be easily completed and interpreted by learners and instructors or need access to technologies that may be permanently available, or only periodically. Thus it needs to be determined whether the assessment is (1) easily used, with minimal or no support; (2) possible to use, but requiring ongoing support; or (3) complex, requiring specialized methods and support.

Technical quality. The assessments may require levels of expertise to administer and score that are beyond the training of many instructors. Technical quality evidence would include not only confirmation that the assessments provide credible information for their intended uses in the environments (e.g., formative or summative), but also that the interpretations of observations and evidence are reliable across instructors and environments. Thus it is important to clarify if technical quality is (1) fully or (2) only partially established.

Connecting Learning Environments and Formative Assessments to Large-Scale Tests

Currently, there are different, often competing, approaches to assessing twenty-first century skills. One approach focuses on assessment *of* technology, such as the International Computer Driving License and technology proficiency tests in some states in the USA. These tests measure the facts and procedures needed to operate common Internet and productivity tools, while the content or the academic or applied problem and context are deliberately chosen to be familiar background knowledge (Venezky and Davis 2002; Crawford and Toyama 2002). The cognitive processes addressed in twenty-first century skills, such as problem solving, communication, collaboration, innovation, and digital citizenship, are not targeted by such tests *of* technology operations.

In a second approach, twenty-first century skills emphasize learning *with* technology by presenting test problems and items that *integrate* measurement of technology operations in terms of strategic use of technology tools to solve problems with subject-matter knowledge and processes, by way of carefully designed sets of tasks and items related to complex academic and real world problems.

In a third approach, the testing is implemented *by* technology. Assessments *by* technology simply use technical infrastructures to deliver and score tests that are designed to measure other content and skills, in subjects such as mathematics and reading. These test designs aim to reduce or eliminate the demands of the technology, treating it as a construct of no relevance. Equivalence of paper-based and technology-based forms is the goal here. Technology-based tests are increasing rapidly in large-scale state, national, and international testing, where technology is being embraced as a way of reducing the costs and logistics of assessment functions, such as test delivery, scoring, and reporting. Technology-based tests typically assume that supporting technology tools such as calculators or word processors are irrelevant to the content constructs being tested and, therefore, are not to be measured separately. Since these types of testing programs seek comparability of paper and online tests, the tests tend to present static stimuli and use traditional constructed-response and selected-response item formats. For the most part, these conventional, online tests remain limited to measuring knowledge and skills that can be easily assessed on paper. Consequently, they do not take advantage of technologies that can measure more complex knowledge structures and the extended inquiry and problem solving included in the twenty-first century ICT skills described in the *Assessment and Teaching of twenty-first Century Skills* project and reported in Chap. 2 (Csapó 2007; Quellmalz and Pellegrino 2009). In short, a technology-delivered and scored test of traditional subjects is not an assessment of twenty-first century ICT skills and should not be taken as one. Twenty-first century skills assessments will not use technology just to support assessment functions such as delivery and scoring, but will also focus on measuring the application of twenty-first century skills while using technology.

Large-scale assessments of twenty-first century skills could provide models of assessments to embed in learning environments, but current large-scale tests do not address the range of twenty-first century skills in ways that would advance knowledge-building environments. In the USA, the new 2012 Framework for Technological Literacy for the National Assessment of Educational Progress sets out three major assessment areas: technology and society, design and systems, and information communication technologies (see naeptech2012.org). Technological literacy in the framework blends understanding of the effects of technology on society, twenty-first century skills, and technology design. The 2012 assessment will present a range of long and short scenario-based tasks designed to assess knowledge and skills in the three areas. In the USA, assessments of twenty-first century skills and technological literacy are required for all students by grade 8. However, state tests or school reports are considered sufficient to meet this requirement, and school reports may be based on teacher reports that, in turn, can be based on questionnaires or rubrics that students use in ICT-supported projects. Most teachers do not have access to classroom assessments of twenty-first century skills, or professional development opportunities to construct their own tests. Moreover, the lack of technical quality of teacher-made and commercially developed classroom assessments is well documented (Wilson and Sloane 2000). Even more of a problem is the lack of clarity for teachers on how to monitor student progress on the development of twenty-first century skills, not only the use of the tools, but ways to think and reason with them. Teachers need formative assessment tools for these purposes.

Concurrent, Embedded and Transformative Assessment of Knowledge Building

In line with the emergence approach as well as the knowledge-creation imperative to continually go beyond what is currently viewed as best practice, we describe new forms of data from classroom environments that make it possible to provide richer, more comprehensive, and more readily available accounts of student performance than are possible through traditional testing. They require new, powerful knowledge-building environments of the sort discussed above.

In the preceding sections we have discussed embedded, formative, and summative assessment; now we add the concepts of concurrent and transformative assessment. Concurrent assessment means that the assessment is available instantaneously. The challenge is effective design of feedback that informs high-level processes as well as more straightforward procedures. Transformative means that the evaluation is not simply an account of past performance, pointing to the next immediate steps, but also provides indication of ways individuals and teams can tackle broader problems and situate their work in relation to that of other team members and teams, within and outside the school walls.

When student discourse is central to the operation of the community, with members contributing to shared, public knowledge spaces, and building on each other's ideas, new forms of assessment make it possible to enrich the community's work and enable concurrent and transformative assessment. The discourse to be analyzed may include online as well as face-to-face interactions, recorded through video or conferencing software and transcribed. Profiles of student work can be generated easily from such data. Even at this early stage, there is a great deal of excitement among the researchers, teachers, and students who have pilot-tested these tools in their classrooms. Teachers and students alike readily see their advantage and generate ideas for improving them.

Data are generated automatically from student discourse and artifacts, and as suggested below, the tools can be used to identify patterns and support continual improvement in practice and student achievement. A substantial part of the challenge in advancing concurrent, embedded, and transformative assessment will be avoiding pitfalls while taking advantage of substantial new opportunities.

Contributions. A contribution tool can provide measures of the number of notes created, the nature of entries (based on keywords, media type, etc.), an overview of the content areas participants worked in, and so forth. Contributions related to a specific problem can be traced, thus making it possible to start investigating individual and group problem solving. The teacher can use the tool during each session or immediately afterward to determine how productive each student has been (e.g., how many notes were read, created, or modified). Such information helps the teacher to direct attention to students who may need more support or instruction, and helps them identify barriers that are preventing students from participating fully in the knowledge-building community. Students can use the tools, if the teacher enables their access, to see where they are in the class distribution (no names are shown).

"Thinking Types" or scaffolds to support twenty-first century skills. Scaffolds can be built on the basis of theory-driven accounts of advanced knowledge processes (see the section on technology to support emergence of new skills). Computer-mediated and customizable scaffold supports 1604 (e.g., "my problem solution," "my theory,") allow teachers and students to use scaffolds and rubrics flexibly and for students to tag their notes according to thinking type (Andrade 2000; Chuy et al. 2009; Law and

Wong 2003; Lai and Law 2006). By identifying the twenty-first century skill they are engaged in (problem solving, theory development, research, decision making, etc.), students become more cognizant of these skills. And once text is tagged, searching by scaffolds makes it easy for students and teachers to find, discuss, and evaluate examples. Formative assessment tools can be used to provide feedback on patterns of use and to help extend students' repertoires.

Use of new media and multiliteracies. Students can contribute notes representing different modalities and media, such as text, images, data tables, graphs, models, video, audio, and so forth. Results suggest that growth in textual and graphical literacy is an important by-product of work in media-rich knowledge-building environments (Sun et al. 2008; Gan et al. 2007).

Vocabulary. A vocabulary tool can provide profiles for individuals and groups, including the rate of new word use, use of selected words from curriculum guidelines (or from any set of words), and so on. It is also easy to look at the growth of vocabulary in comparison to external measures or benchmarks, such as grade-level lists. Thus teachers can determine if important concepts are entering the students' productive vocabularies, the extent of their use of words at or above grade level, their growth in vocabulary based on terms at different levels in the curriculum guidelines, and so on. Information about the complexity and quality of notes can also give the teacher direction as to the type of instruction the class may need. Early, informal use of these vocabulary tools suggests that students enjoy seeing the growth in their vocabulary, and begin to experiment with new words that have been used by others in the class.

Writing. Measures of writing start with basic indicators (e.g., total and unique words, mean sentence length). There are many sophisticated tools already developed, and open-source arrangements will make it increasingly easy to link discourse and writing environments.

Meta-perspectives. A brainstorming tool (Nunes et al. 2003) can be used to foster students' metacognitive thinking about specific skills and support students in the exercise of creativity, leadership, and collaboration. Tools can also be built to allow students to tag notes containing questions asked but not answered, claims made with no evidence, etc. Once tagged, visualization tools can bring to the forefront of the knowledge space ideas needing extra work.

Semantic analysis. This tool makes it possible to work in many and flexible ways with the meaning of the discourse. A semantic-overlap facility extracts key words or phrases from user-selected texts and shows overlapping terms. One application of this tool is to examine overlapping terms between a participant's discourse and discourse generated by experts or in curriculum guidelines. Other applications include examination of overlapping terms between texts of two participants or between a student text and an assigned reading. A semantic field visualization provides graphical displays of the overlapping terms by employing techniques from latent semantic analysis (Teplovs 2008). For example, a benchmark can be identified (an encyclopedia

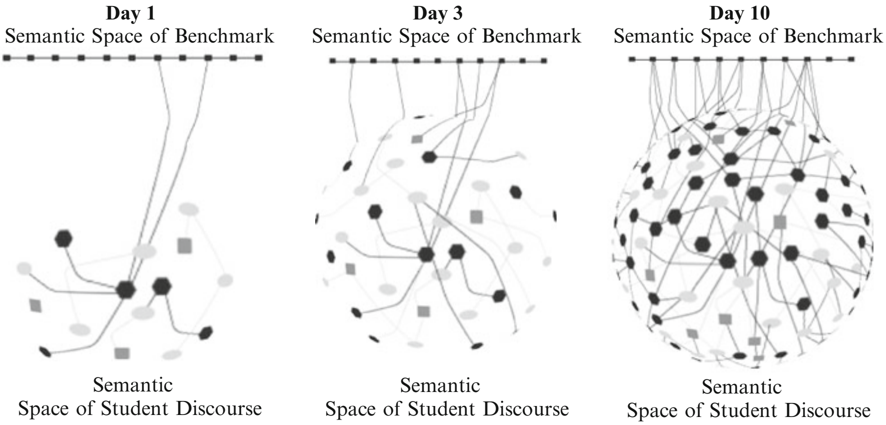

Day 1
Semantic Space of Benchmark
Semantic
Space of Student Discourse
Day 3
Semantic Space of Benchmark
Semantic
Space of Student Discourse
Day 10
Semantic Space of Benchmark
Semantic
Space of Student Discourse

Fig. 5.4 Semantic field visualization of a classroom over 10 days (Adapted from Teplovs 2008)

entry, a curriculum guideline or standard, etc.). The tool can show the overlap between the students' discourse and the benchmark over successive days, as the visualization in Fig. 5.4 suggests.

Social network analysis. Social network analysis tools display the social relationships among participants based on patterns of behavior (e.g., who read/referenced/built on whose note). A social network analysis tool can help teachers to better understand who the central participants are in the knowledge-building discourse and to see whether existing social relationships are limiting the community's work or influencing it positively. The tool draws the teacher's attention to children who are on the periphery and makes it more likely that these children will receive the support they may need to be more integral to the work of the class.

Increasing levels of responsibility for advancing collective knowledge is facilitated when student contributions to classroom work are represented in a communal knowledge space. Below are graphics generated from the social network analysis tool to give some sense of how it is possible to uncover classroom practices associated with advances in student performance—practices that would be impossible to uncover without use of communal discourse spaces. The work reported in Fig. 5.5 (Zhang et al. 2007, 2009) is from a grade 4 classroom studying optics. The teacher and students worked together to create classroom practices conducive to sustained knowledge building. Social network analysis and independently generated qualitative analyses were used to assess online participatory patterns and knowledge advances, focusing on indicators of collective cognitive responsibility.

The social network graphs generated by the *social network analysis tool* indicate increasingly effective procedures for advancing student knowledge corresponding to the following social organizations: (a) year 1—fixed, small-groups; (b) year 2—interactive small groups working together throughout their knowledge work; and (c) year 3—opportunistic collaboration, with small teams forming and disbanding under the

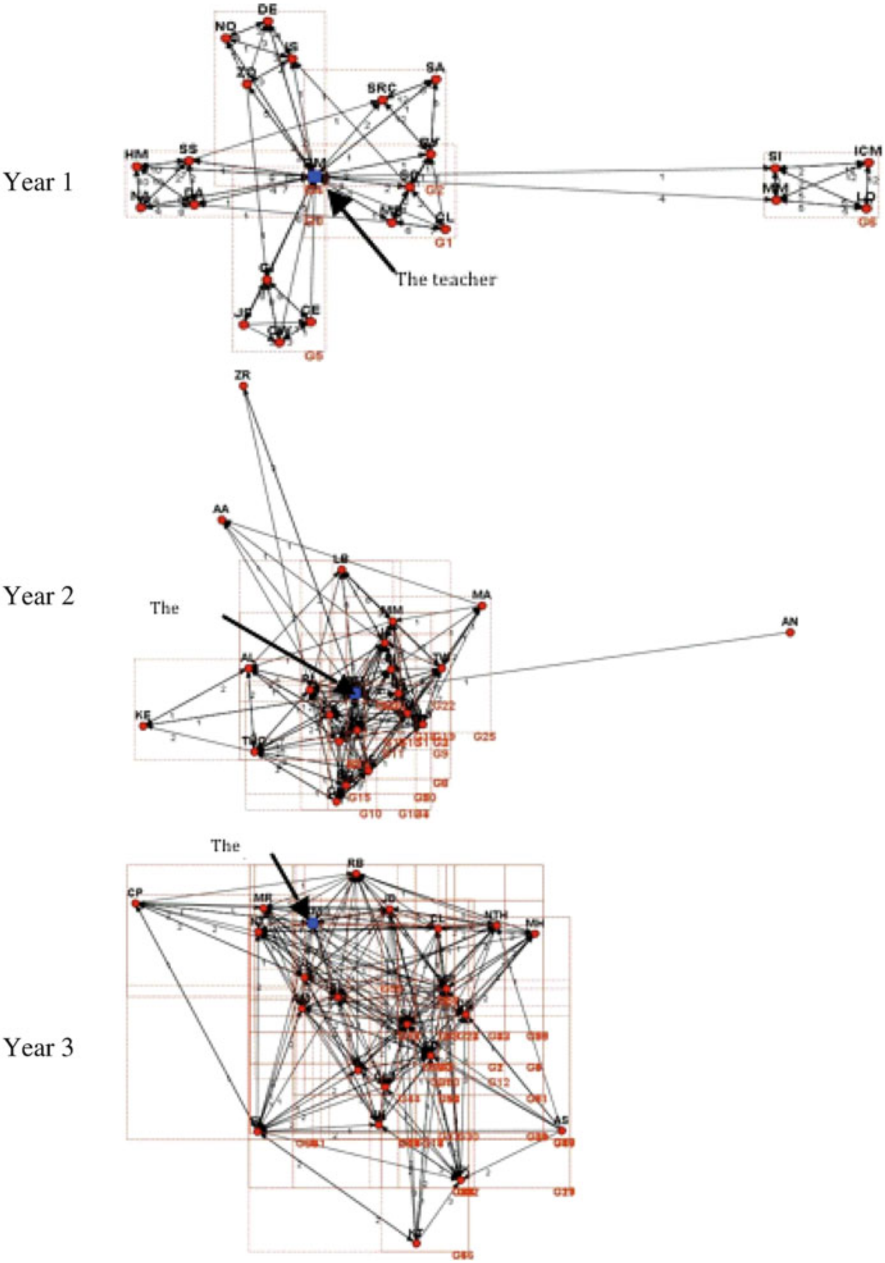


Fig. 5.5 The emergent process of knowledge building over 3 years (This 3-year account, from the perspective of the social network analysis tool, is described in detail in Zhang et al. 2009)

volition of community members, based on emergent goals that arose as they addressed their shared, top-level goal of refining their knowledge of optics. The third-year model maps most directly onto the organic and distributed social structure in real-world knowledge-creating organizations. Among the three designs, the opportunistic-collaboration model resulted in the highest level of collective cognitive responsibility, knowledge advances, and dynamic diffusion of information. This 3-year account, as shown from the perspective of the social network analysis tool, is shown in the following figure – Fig. 5.5 (see Zhang et al. 2009 for details).

In these graphs a node represents a group member. A line between two nodes denotes a note linking relation between two members, indicating that one member has built on or referred to a note by the other. The direction and frequency of such connections are represented by the arrow and value on the line. The more information flow a member carries, the more centrally he/she is displayed in a network. Tools such as those presented above allow teachers and students to visualize their work in new ways. They can be applied to discourse on any topic, at the group as well as individual level. There are endless possibilities for reconstructing knowledge spaces to bring different issues and concerns into perspective and to show change over time. This work is in its infancy and Web 2.0/3.0 developments will greatly enhance it.

Assessment, Open Knowledge Resources, and Development of Knowledge Building

The need for developmental frameworks, definitions, and models can be seen throughout the *Assessment and Teaching of Twenty-First Century Skills project*. This is evident in the discussion of frameworks (Chap. 2), the argument for the need to identify learning progressions to describe pathways that learners are likely to follow toward the mastery of a domain (Chap. 3), and the discussion of item development (Chap. 4). We hope to contribute to these efforts through identifying developmental progressions grounded in the theory and practices of knowledge-creating organizations. We argue that all citizens should have the opportunity to participate in knowledge-building environments that fully integrate twenty-first century skills and move them along the developmental trajectories set out earlier in Table 5.2. The tools we describe above can help accomplish this by charting progress and addressing design principles in new ways.

Design principles for knowledge-building environments include: (a) empowering users and transferring greater levels of agency and collective responsibility to them; (b) viewing assessment as integral to efforts to advance knowledge and identify problems as work proceeds; (c) enabling users to customize tools and request changes so that the environments are powerful enough to be embedded in the day-to-day workings of the organization; (d) supporting the community in self-directed rigorous assessment so that there is opportunity for the community's work to exceed, rather than simply meet expectations of external assessors; (e) incorporating standards and benchmarks into the process so that they are entered into the public workspace in digitized form and become objects of discourse that can be annotated, built on, linked to ongoing work, and risen above; (f) supporting inclusive design, so there is a way in for all participants; this challenge brings with it special technological challenges (Trevinarus 1994, 2002); (g) providing a public design space to support discourse around all media (graphics, video, audio, text, etc.) with links to all knowledge-rich and domain-specific learning environments; and (h) encouraging openness in knowledge work. Once these requirements are met, participants are engaged with ICT in meaningful, interactive contexts, with reading and writing part of their expressive work across all areas of the school curriculum. They can then make extensive use of the forms of support that prove so helpful in knowledge-creating organizations—connections with other committed knowledge workers and world-class knowledge resources.

Combining ICT-enabled discourse environments and open resources sets the stage for breakthroughs in charting and enhancing development in knowledge-building environments. For example, student discourse environments can be linked to powerful simulation, tutorial, intelligent tutoring, and other domain-specific tools (Quellmalz and Haertel 2008; Tucker (2009); http://www.ascd.org/publications/educational_leadership/nov09/vol67/num03/The_Next_Generation_of_Testing.aspx; http://oli.web.cmu.edu/openlearning/initiative). It is then possible to combine the benefits of these different tools and promote interactions surrounding their use. As explained in The Open Learning Initiative, Carnegie Mellon University, it is possible to build assessment "into every instructional activity and use the data from those embedded assessments to drive powerful feedback loops for continuous evaluation and improvement." Assessments from these tutorials, simulations, games, etc., can complement those described in the section on open-source software and programming interfaces and, combined with interoperability of applications, allow us to further break down the barriers between various environments and assessments that have traditionally been separate and disconnected, so as to search and compile information across them. Open resources make it possible to assemble information on learning progressions, benchmarks, and learning modules. Curriki is an example of a web site where the community shares and collaborates on free and open-source curricula (http://www.curriki.org/). Creative Commons licenses further expand access to information to be shared and built upon, bringing an expanded concept of intellectual property.

These open resources, combined with data from discourse environments, make it possible to build student portfolios, based on classroom work and all the web-accessible information created from in- or out-of-school uses of simulations, games, etc., across topics and applications (dealing with ethical issues presents a different, significant challenge). Extended student portfolios will allow us to chart student progress in relation to various and changing developmental benchmarks, as well as to foster development through formative feedback. For example, "nearest neighbor" searches, based on student semantic spaces, can identify other people, in the same class or globally, as well as local or global resources, working with similar content. Connections can then be made, just in time, any time, to meet both teacher and student needs. This support can help the class as a whole to operate as a twenty-first century organization, as well as supporting individual student achievement.

We envision worldwide teams of users (Katz et al. 2009) and developers taking advantage of new data-mining possibilities, intelligent web applications, semantic analysis, machine learning, natural language processing, and other new developments to advance the state of the art in education.

Technology to Support Emergence of New Competencies

Two recent books discuss in depth the effects that new technologies can have in shifting education on to a new basis for the twenty-first century. One is *Rethinking Education in the Age of Technology: The Digital Revolution and Schooling in America* (Collins and Halverson 2009). Collins and Halverson argue that new technologies create learning opportunities that challenge traditional schools. They envision a future in which technology enables people of all ages to pursue learning on their own terms. Figure 5.3 above indicates that more time by far is spent in out-of-school contexts, across the entire lifespan. If these become primary contexts for learning, tasks designed especially for school will pale by comparison in their impact on education. The second book is *The World Is Open: How Web Technology Is Revolutionizing Education* (Bonk 2009). Bonk explains ways in which technologies have opened up the education world to anyone, anywhere. He discusses trends such as web searching, open courseware, real-time mobility, portals, and so forth that will impact learning in the twenty-first century. These technologies are not envisaged as a cafeteria line for students to proceed along and pick and choose (which, unfortunately, seems to have been the formative concept in many instructional support systems); instead, they are envisaged as constituting an environment supportive of a more fully engaged community of learners, more open to the world's cognitive and emotional riches.

These ideas are in line with our earlier discussions of the emergence of new competencies and open resources. Rather than simply extrapolating from existing goals or expert-identified objectives, new goals can emerge from the capacities that students demonstrate in supportive environments—such as the capacities for proportional reasoning and theory building revealed in the examples cited. Both these experimental approaches have, in fact, made use of computer-supported knowledge-building environments that provide support for the creation of public knowledge (Moss and Beatty 2006; Messina and Reeve 2006). Among the technical affordances serving this purpose are "thinking types" or scaffolds, described above, "rise-above" notes that serve the purposes of synthesis and the creation of higher-order representations of ideas, and graphical backgrounds for creating multiple representations and organizing ideas (Scardamalia and Bereiter 2006).

In the theory-building work elaborated above, scaffolds supported theory building. The "theory supports" included the following phrases: "My theory," "I need to understand," "Evidence for my theory," "Putting our knowledge together," "A better theory." To use these scaffolds, students simply need to click on one of these phrases, arrayed on a panel to the left of their writing space, and a text field containing the phrase is copied into their text at the appropriate point. Text added by the student is automatically tagged according to the scaffold name. This simple support has increased the use of these phrases in student writing and, results suggest, has enhanced the high-level knowledge processes they represent. In the Knowledge Forum environment, used in the theory-building example, scaffolds are customizable, so these discourse supports can easily be changed to fit any twenty-first century goal. (They can also be used after the fact, to mark up text already written.) These scaffolds foster metacognitive awareness, as students use them to characterize their discourse. The scaffold supports also serve as search parameters, further encouraging their use and allowing students and teachers easily to search their communal knowledge space so as to determine what different theories there are in the database, what evidence is used to defend them, the nature of theories that are considered to be improvements on earlier theories, and so forth. And it is quite easy with these tagged "thinking types" to build formative assessments to enhance student development. For example, it is possible to create profiles of student or group activity, to find whether students and the class are generating lots of theories but providing no evidence—or perhaps they are providing evidence but cannot put their ideas together to generate an improved theory. Patterns of use make it possible to detect underrepresented knowledge processes and to inform and advance such work.

An important role for technology is to support individuals in constructive contributions to the group. The scaffolds help. At the group level the essential question is: Has the public knowledge shared by a group progressed—to what extent has this knowledge emerged from a group process as opposed to being merely an aggregation of individual products? Web 3.0 "semantic web" developments treat ideas or meanings rather than simply words as the units of primary interest. Some educational evaluation tools have already taken advantage of these advances (Teplovs 2008) and we can look forward to further developments that align more powerful web technology with educational needs for working in a knowledge-creating culture. We elaborate on these ideas in the section on technological and methodological advances to support the development of twenty-first century skills.

Although findings from the emergence approach are limited, they suggest that students demonstrate advances across a broad range of twenty-first century skills (Chuy et al. 2009; Gan et al. 2007; Sun et al. 2008, 2010), and that an emergence approach may contribute genuinely new discoveries to inform large-scale assessment. Positive results of an emergence approach also suggest that defining and operationalizing twenty-first century skills one-by-one, while important for measurement purposes, may not be the best basis for designing educational activity.

As technology blurs the line between in- and out-of-school contexts, and knowledge becomes a social product situated in open worlds, the need for environments and formative assessment that span educational contexts and support "community knowledge" and group or "collective intelligence" will become increasingly important.

Necessary Research

This section identifies important areas of research and development related to the overall goal of developing new assessments and environments for twenty-first century knowledge building. We start with research and development to improve formative assessments in current learning environments and then move on to studies and advances in formative assessment likely to transform schools into the image of knowledge-creating organizations.

Analysis of Twenty-First Century Skills in Current Learning Environments

A research program on reforming the assessment of twenty-first skills would benefit from greater understanding of twenty-first century skills as represented in current learning environments. Projects could be selected to represent various learning environments, and assessments would focus on twenty-first century skills frameworks and developmental trajectories. We anticipate that all of the learning environments will show limits in the extent to which they address twenty-first century skills, and this analysis could provide important information for evidence-centered initiatives to promote these skills.

The second phase of the study would analyze the technical quality of the projects' assessments and their utility for providing formative evidence during instruction. Using the evidence-centered design framework, we anticipate that there will be weak links between assessments of twenty-first century skills, learning tasks used to elicit those skills, and the evidence that teachers and students can use to understand development of the skills.

A third phase of the study would involve the creation of evidence-centered classroom assessment systems with representative projects to address all or many of the twenty-first century skills. Technical quality data would be collected about their reliability and validity for classroom formative purposes. In addition, the designs of the formative twenty-first century assessments would be linked to the more compressed, constrained designs of the large-scale, summative twenty-first century assessment tasks being designed by all ATC21S working groups. Classroom formative assessments would be embedded in the learning activities, provide evidence of ongoing learning processes related to twenty-first century skills, such as problem solving, collaboration, and communication, and would provide rich, deep, frequent streams of evidence to be used by learners and instructors during their learning activities to monitor and support their progress. For example, in domain-centered learning environments, such rich, embedded formative assessment would be made possible by digital capture of student processes during domain-specific learning activities such as information research, use of simulations, and network analyses. The study would examine the formative utility and technical quality of the assessments and the value they had added to interim benchmark summative assessments and to even more distal large-scale state, national, and international assessments. The research on the design of quality formative assessments for the full range of twenty-first century skills that could be embedded in projects in each of the different learning environments would serve as models for reforming and transforming twenty-first century formative assessments in learning environments.

Social and Technological Innovations for An Inclusive Knowledge-Building Society

The goals currently being promoted for twenty-first century skill development are, as previously noted, based mainly on expert and stakeholder analysis of goals. In this section we propose design experiments that complement this top-down approach to goal identification with a bottom-up approach based on the capacities, limitations, and problems that learners reveal when they are actually engaged in knowledge-creating work. The first step in mounting such research is to identify or establish schools able to operate as knowledge-creating organizations—given, as Laferrière and Gervais (2008) suggest—that at this point it may be difficult to locate schools able to take on such work. The proposed research has the dual purpose of (a) discovering previously unrecognized skill goals and (b) developing ways of assessing these emergent skills through minimally intrusive instruments.

Sites thus engaged, willing to take on an ambitious new research agenda, and equipped with appropriate technology, could then support a broad-based research and development effort aimed at addressing questions related to knowledge practices and outcomes. At a policy level we would begin to collect data and evidence to address issues that are dividing educators. For example, many educators favor those curriculum procedures and processes that are well defined and have a step-by-step character—but knowledge creation is not an orderly step-by-step process. Knowledge creators go where their ideas take them. How can the challenge of engaging students in more self-directed and creative work with ideas be reconciled with the classroom routines and activity structures that many educators feel to be essential for teachers, students, and curriculum coverage? How does self-organization, an important component of knowledge creation, actually combine with intentional development of ideas at the process level? How are promising ideas worthy of further development sorted out from the large pool of ideas students often generate? How can "pooling of ignorance" be avoided?

"Pooling of ignorance" is a problem that looms large in discussions about open discourse environments for naïve learners. Although "making thinking visible" is one of the advantages claimed for constructivist computer environments, it can increase the chances of "pooling ignorance" and spreading "wrong" ideas. Teachers, accordingly, are tempted to exert editorial control over what ideas get made public in student inquiry; and students, for their part, may learn that it's better to put forward authoritative ideas, rather than their own. Research is needed, first to determine whether "pooling ignorance" is a real or only an imagined problem, and second—if it does prove to be real—to carry out design research to find a constructive way to deal with this dilemma.

Concurrent, embedded, and transformative assessments need to be geared to demonstrations of new ways around old problems. We can then collectively test the notion that formative assessments, built into the dynamics of the community, will allow for a level of self-correction and a focus on high-level goals unparalleled in most educational contexts.

Challenges Related to Complex Interventions

Brown (1992), Collins et al. (2004) and Frederiksen and Collins (1989) discuss theoretical and methodological challenges in creating complex interventions and the problems of narrow measures. They stress the need for design experiments as a way to carry out formative research for testing and refining educational designs based on theoretical principles derived from prior research. It is an approach of "progressive refinement." As Collins et al. (2004) explain, design experimentation

> involves putting a first version of a design into the world to see how it works. Then, the design is constantly revised based on experience… Because design experiments are set in learning environments, there are many variables that cannot be controlled. Instead, design researchers try to optimize as much of the design as possible and to observe carefully how the different elements are working out. (p.18)

Chapter 3 raises a number of methodological issues regarding assessment of twenty-first century skills. The proposed research could contribute to progress on each of the issues raised there: (a) *Distinguishing the role of context from that of the underlying cognitive construct*—the experiment would allow us to find examples of the construct across different national and domain contexts; (b) *new types of items that are enabled by computers and networks*—the network we propose would implement new designs and explore uses of new item types; (c) *new technologies and new ways of thinking to gain more information from the classroom without overwhelming the classroom with more assessments*—we propose to engage a network of international, multilingual, cross-domain centers to explore issues and determine how concurrent, embedded, and transformative assessments might begin to save teachers' time; (d) *right mix of crowd wisdom and traditional validity*—"crowd wisdom" and traditional procedures can easily be combined in the environments we propose; (e) *information and data availability and usefulness*—we can directly explore what it takes to translate data into feedback to drive knowledge advancement; and (f) *assessments for twenty-first century skills that are activators of students' own learning*—through the use of scaffolds, adaptive recommender systems, stealth assessments, visualizations, and so on, we can explore assessments that facilitate students' own learning.

Specific Investigations Within the Emergent Competencies Framework

We propose that an international network of pilot sites be established, both to cooperate in the multifaceted design research described below and to collaborate with local researchers in creating and testing new designs tailored to their own conditions and needs. A given site may collaborate in all or a subset of the specific investigations, but in any event the data they produce will be available for addressing the full range of research questions that arise within the network. The following, therefore, should be regarded as an initial specification, subject to modification and expansion.

Charting developmental pathways with respect to twenty-first century skills. As indicated in the sections on embedded assessment and technology to support the emergence of new skills, computer-based scaffolds can be used to support the development of twenty-first century skills and formative assessments related to their use. An intensive program of research to develop each skill would allow us to determine what students at various ages are able and not able to do related to various twenty-first century skills, with and without supports for knowledge creation. We would then be in a better position to elaborate the developmental progressions set out in Table 5.2.

Demonstrating that knowledge-building pedagogy saves educational time rather than adding additional, separate skills to an already crowded curriculum. Currently, learning basic skills and creating new knowledge are thought by many to be competitors for school time. In knowledge-building environments, students are reading, writing, producing varied media forms, and using mathematics to solve problems—not as isolated curriculum goals but through meaningful interactions aimed at advancing their understanding in all areas of the curriculum. Rather than treating literacy as a prerequisite for knowledge work, it becomes possible to treat knowledge work as the preferred medium for developing multiliteracies. Early results indicate that there are gains in subject-matter learning, multiliteracies, and a broad range of twenty-first century skills. These results need to be replicated and extended.

Testing new technologies, methods, and generalization effects. The international network of pilot sites would serve as a test bed for new tools and formative assessments. In line with replication studies, research reported by Williams (2009) suggests that effective collaboration accelerates attainments in other areas. This "generalization effect" fits with our claim that, although defining and operationalizing twenty-first century skills one-by-one may be important for measurement purposes, educational activities will be better shaped by a more global conception of collaborative work with complex goals. Accordingly, we propose to study relationships between work in targeted areas and then expand into areas not targeted. For instance, we may develop measures of collaborative problem solving, our target skill, and then examine its relationship with collaborative learning, communication, and other twenty-first century skills. We would at the same time measure outcomes on an appropriate achievement variable relevant to the subject matter of the target skill. Thus we would test generalization effects related to the overall goal of educating students for a knowledge-creating culture.

Creating inclusive designs for knowledge building. It is important to find ways for all students to contribute to the community knowledge space, and to chart advances for each individual as well as for the group as a whole. Students can enter into the discourse through their favorite medium (text, graphics, video, audio notes) and perspective, which should help. Results show advances for both boys and girls, rather than the traditional finding in which *girls* outperform *bo*ys in literacy skills. This suggests that boys lag in traditional literacy programs because they are not rewarding or engaging, whereas progressive inquiry both rewards and engages. New designs to support students with disabilities will be an essential addition to environments to support inclusive knowledge building

Exploring multilingual, multiliteracy, multicultural issues. Our proposed research would engage international teams; thus it would be possible to explore the use of multilingual spaces and possibilities for creating multicultural environments. More generally, the proposed research would make it possible to explore issues of a knowledge-building society that can only be addressed through a global enterprise.

Administering common tests and questionnaires. While there is currently evidence that high-level knowledge work of the sort identified in Table 5.1 for knowledge-creating organizations can be integrated with schooling, starting no later than the middle elementary grades (Zhang et al. 2009), data are needed to support the claim that knowledge building is feasible across a broad range of ages, SES contexts, teachers, and so forth, and that students are more motivated in knowledge-building environments than in traditional environments. To maximize knowledge gains from separate experiments, it will be important to standardize on assessment tools, instruments, and data formats. Through directed assessment efforts, it will be possible to identify parameters and practices that enable knowledge building (Law et al. 2002).

Identifying practices that can be incorporated into classrooms consistent with those in knowledge-creating organizations. By embedding practices from knowledge-creating organizations into classrooms, we can begin to determine what is required to enable schools to operate as knowledge-creating organizations and to design professional development to foster such practices. Data on classroom processes should also allow us to refine the developmental trajectory set out in Table 5.2, and build assessments for charting advances at the individual, group, and environment levels.

Demonstrating how a broader systems perspective might inform large-scale, on-demand, summative assessment. We have discussed the distinction between a "working-backward" and "emergence" approach to advance twenty-first century skills and connections between knowledge-building environments, formative assessments, and large-scale assessment. Within the emergence approach, connections between student work and formative and summative assessment can be enriched in important ways. For example, as described above, scaffolds can be built into the environments to encourage students to tag "thinking types." As a result, thinking is made explicit and analytic tools can then be used to assess patterns and help to inform next steps. With students more knowledgeably and intentionally connected to the achievement of the outcomes to be assessed, they can become more active players in the process. In addition to intentionally working to increase their understanding relative to various learning progressions and benchmarks, they are positioned to comment on these and exceed them. As in knowledge-creating organizations, participants are aware of the standards to be exceeded. As an example, toward the end of student work in a unit of study, a teacher, published relevant curriculum standards in the students' electronic workspaces so they could comment on these standards and on how their work stood up in light of them. The students noted many ways in which their work addressed the standards, and also important advances they had made that were not represented in the standards. We daresay that productive dialogues between those tested and those designing tests could prove valuable to both parties. Semantic analysis tools open up additional possibilities for an emergence framework to inform large-scale assessments. It is possible to create the "benchmark corpus" (the semantic field from any desired compilation of curriculum or assessment material), the "student corpus" (the semantic field from any desired compilation of student-generated texts such as the first third of their entries in a domain versus the last third), and the "class corpus" (the semantic field from all members of the class, first third versus last third), and so forth. Semantic analysis and other data-mining techniques can then be used to track and inform progress, with indication of semantic spaces underrepresented in either the student or benchmark corpus, and changes over time.

Classroom discourse, captured in the form of extensive e-portfolios, can be used to predict performance on large-scale summative assessments and then, through formative feedback, increase student performance. Thus results can be tied back to performance evaluations and support continual improvement. Teachers, students, and parents all benefit, as they can easily and quickly monitor growth to inform progress. This opens the possibility for unprecedented levels of accountability and progress.

Technological and Methodological Advances to Support Skills Development

Technological advances, especially those associated with Web 2.0 and Web 3.0 developments, provide many new opportunities for interoperability of environments for developing domain knowledge and supporting student discourse in those domains. Through coherent media-rich online environments, it is possible to bring ideas to the center and support concurrent, embedded, and transformative assessment. As indicated above, it is now possible to build a broad range of formative assessments that will enrich classroom work greatly.

A key characteristic of Web 2.0 is that users are no longer merely consumers of information but rather active creators of information that is widely accessible by others. The concomitant emergence of online communities, such as MySpace, LinkedIn, Flickr, and Facebook, has led, ironically and yet unsurprisingly, to a focus on individuals and their roles in these communities as reflected, for example, in the practice of counting "friends" to determine connectedness. There has been considerable interest in characterizing the nature of social networks, with social network analysis employed to detect patterns of social interactions in large communities. Web 3.0 designs represent a significant shift to encoding semantic information in ways that make it possible for computers to deduce relationships among pieces of information. In a Web 3.0 world the relationships and dynamics among ideas are at least as important as those among users. As a way of understanding such relationships, we can develop an analogue of social network analysis—*idea* network analysis. This is especially important for knowledge-building environments where the concern is social interactions that enable idea improvement (see Teplovs 2008). Idea network analysis offers a means of describing relationships among ideas, much as social network analysis describes the relationships among actors. Visualizations of idea networks, with related metrics such as network density, will allow us to characterize changes in social patterns and ideas over time. The demanding conceptual and research challenge, therefore, is to understand and support the social dynamics that lead to knowledge advancement.

Through additional design work, aimed at integrating discourse environments, online knowledge resources, and formative and summative assessments, we can greatly extend where and how learning might occur and be assessed. By tracking the semantics of participant discourses, online curriculum material, test items, texts of experts in the field, and so on, we can map one discourse or corpus onto another and track the growth of ideas. With collaborative online discourse integral to the operation of knowledge-building communities, we can further enhance formative assessments so as to encourage participants to seek new learning opportunities and a broader range of experts.

Effectively designed environments should make it possible to develop communication, collaboration (teamwork), information literacy, critical thinking, ICT literacy, and so forth in parallel—a reflection of how things work in knowledge-creating organizations.

Annex: Knowledge-Building Analytic Framework

Template for Analyzing Environments and Assessments

1. DESCRIBE AN ENVIRONMENT AND/OR ASSESSMENT AS IT CURRENTLY EXISTS. (*Use as much space as you need)*

2. INDICATE WHETHER THE EXAMPLE FITS PRIMARILY INTO AN ADDITIVE OR TRANSFORMATIVE MODEL OF SCHOOL REFORM. TO PROVIDE THIS EVALUATION, YOU SIMPLY NEED TO ASSIGN A SCORE FROM 1 (definitely additive) to 10 (definitely transformative), AND PROVIDE A BRIEF RATIONALE. NOTE: Score = 1 (the goal is *additive* if the environment, or assessment presented is designed to add a task or activity to school work that remains little changed in overall structure, other than through the addition of this new task, project, environment or assessment*); Score = 10 (the goal is transformative if the environments or assessment alters conditions of schooling in a substantial way, so students become enculturated into a knowledge-creating organization that is supported by a knowledge-building environment integral to the operation of the community).*

   SCORE _______

RATIONALE FOR SCORE: (*Use as much space as you need)*

3. PLEASE USE THE FOLLOWING EVALUATION FORM TO ASSESS THE CHARACTERISTICS OF THE ENVIRONMENT AND/OR ASSESSMENT IN ITS CURRENT FORM

Twenty-first century skill (from Chap. 2)	Characteristics of knowledge-creating organizations: a continuum that maps onto twenty-first century skills		
	1	5	10

Creativity and innovation

SCORE FROM 1 (internalize given information; beliefs/actions based on the assumption that someone else has the answer or knows the truth) to 10 (work on unsolved problems; generate theories and models, take risks, etc; pursue promising ideas and plans)

SCORE_______

RATIONALE FOR YOUR SCORE:
(*Use as much space as you need*)

DO YOU SEE A WAY TO IMPROVE YOUR ENVIRONMENT OR ASSESSMENT ALONG THIS DIMENSION? IF SO, PLEASE PROVIDE A BRIEF ACCOUNT OF HOW YOU MIGHT DO THAT, OR HOW THE IDEAS IN THIS WORKING PAPER MIGHT HELP.
(*Use as much space as you need*)

Communication

SCORE FROM 1 (social chitchat; discourse that aims to get everyone to some predetermined point; limited context for peer-to-peer or extended interactions) to 10 (knowledge building/progressive discourse aimed at advancing the state of the field; discourse to achieve a more inclusive, higher-order analysis; open community knowledge spaces encourage peer-to-peer and extended interactions)

SCORE_______

RATIONALE FOR YOUR SCORE:
(*Use as much space as you need*)

DO YOU SEE A WAY TO IMPROVE YOUR ENVIRONMENT OR ASSESSMENT ALONG THIS DIMENSION? IF SO, PLEASE PROVIDE A BRIEF ACCOUNT OF HOW YOU MIGHT DO THAT, OR HOW THE IDEAS IN THIS WORKING PAPER MIGHT HELP.
(*Use as much space as you need*)

Collaboration/teamwork

SCORE FROM 1 (small group work—divided responsibility to create a finished product; the whole is the sum of its parts, not greater than that sum) to 10 (collective or shared intelligence emerges from collaboration and competition of many individuals and aims to enhance the social pool of existing knowledge. Team members aim to achieve a focus and threshold for productive interaction and work with networked ICT. Advances in community knowledge are prized, over-and-above individual success, while enabling each participant to contribute to that success)

SCORE_______

RATIONALE FOR YOUR SCORE:
(*Use as much space as you need*)

DO YOU SEE A WAY TO IMPROVE YOUR ENVIRONMENT OR ASSESSMENT ALONG THIS DIMENSION? IF SO, PLEASE PROVIDE A BRIEF ACCOUNT OF HOW YOU MIGHT DO THAT, OR HOW THE IDEAS IN THIS WORKING PAPER MIGHT HELP.
(*Use as much space as you need*)

Information Literacy/research

SCORE FROM 1 (inquiry: question-answer, through finding and compiling information; variable testing research) to 10 (going beyond given information; constructive use of and contribution to knowledge resources to identify and expand the social pool of improvable ideas, with research integral to efforts to advance knowledge resources and information)

SCORE_______

RATIONALE FOR YOUR SCORE:
(*Use as much space as you need*)

DO YOU SEE A WAY TO IMPROVE YOUR ENVIRONMENT OR ASSESSMENT ALONG THIS DIMENSION? IF SO, PLEASE PROVIDE A BRIEF ACCOUNT OF HOW YOU MIGHT DO THAT, OR HOW THE IDEAS IN THIS WORKING PAPER MIGHT HELP.
(*Use as much space as you need*)

Critical thinking, problem solving and decision-making

SCORE FROM 1 (meaningful activities are designed by the director/teacher/curriculum designer; learners work on predetermined tasks set by others.) to 10 (high-level thinking skills exercised in the course of authentic knowledge work; the bar for accomplishments is continually raised through self-initiated problem finding and attunement to promising ideas; participants are engaged in complex problems and systems thinking)

SCORE_______

RATIONALE FOR YOUR SCORE:
(*Use as much space as you need*)

DO YOU SEE A WAY TO IMPROVE YOUR ENVIRONMENT OR ASSESSMENT ALONG THIS DIMENSION? IF SO, PLEASE PROVIDE A BRIEF ACCOUNT OF HOW YOU MIGHT DO THAT, OR HOW THE IDEAS IN THIS WORKING PAPER MIGHT HELP.
(*Use as much space as you need*)

Citizenship—local and global

SCORE FROM 1 (support of organization and community behavioral norms; "doing one's best"; personal rights) to 10 (citizens feel part of a knowledge-creating civilization and aim to contribute to a global enterprise; team members value diverse perspectives, build shared, interconnected knowledge spanning formal and informal settings, exercise leadership, and support inclusive rights)

SCORE_______

RATIONALE FOR YOUR SCORE:
(*Use as much space as you need*)

DO YOU SEE A WAY TO IMPROVE YOUR ENVIRONMENT OR ASSESSMENT ALONG THIS DIMENSION? IF SO, PLEASE PROVIDE A BRIEF ACCOUNT OF HOW YOU MIGHT DO THAT, OR HOW THE IDEAS IN THIS WORKING PAPER MIGHT HELP.
(*Use as much space as you need*)

ICT literacy

SCORE FROM 1 (familiarity with and ability to use common applications and web resources and facilities) to 10 (ICT integrated into the daily workings of the organization; shared community spaces built and continually improved by participants, with connection to organizations and resources worldwide)

SCORE_______

RATIONALE FOR YOUR SCORE:
(*Use as much space as you need*)

DO YOU SEE A WAY TO IMPROVE YOUR ENVIRONMENT OR ASSESSMENT ALONG THIS DIMENSION? IF SO, PLEASE PROVIDE A BRIEF ACCOUNT OF HOW YOU MIGHT DO THAT, OR HOW THE IDEAS IN THIS WORKING PAPER MIGHT HELP.
(*Use as much space as you need*)

Life and career skills

SCORE FROM 1 (personal career goals consistent with individual characteristics; realistic assessment of requirements and probabilities of achieving career goals) to 10 (engagement in continuous, "lifelong" and "life-wide" learning opportunities; self-identification as a knowledge creator, regardless of life circumstance or context)

SCORE_______

RATIONALE FOR YOUR SCORE:
(*Use as much space as you need*)

DO YOU SEE A WAY TO IMPROVE YOUR ENVIRONMENT OR ASSESSMENT ALONG THIS DIMENSION? IF SO, PLEASE PROVIDE A BRIEF ACCOUNT OF HOW YOU MIGHT DO THAT, OR HOW THE IDEAS IN THIS WORKING PAPER MIGHT HELP.
(*Use as much space as you need*)

Learning to learn/
meta-cognition

SCORE FROM 1 (students and workers provide input to the organization, but the high-level processes are under the control of someone else) to 10 (students and workers are able to take charge at the highest, executive levels; assessment is integral to the operation of the organization, requiring social as well as individual metacognition)

SCORE_______

RATIONALE FOR YOUR SCORE:
(*Use as much space as you need*)

DO YOU SEE A WAY TO IMPROVE YOUR ENVIRONMENT OR ASSESSMENT ALONG THIS DIMENSION? IF SO, PLEASE PROVIDE A BRIEF ACCOUNT OF HOW YOU MIGHT DO THAT, OR HOW THE IDEAS IN THIS WORKING PAPER MIGHT HELP.
(*Use as much space as you need*)

Personal and social responsibility—incl. cultural competence

SCORE FROM 1 (individual responsibility; local context) to 10 (team members build on and improve the knowledge assets of the community as a whole, with appreciation of cultural dynamics that will allow the ideas to be used and improved to serve and benefit a multicultural, multilingual, changing society)

SCORE_______

RATIONALE FOR YOUR SCORE:
(*Use as much space as you need*)

DO YOU SEE A WAY TO IMPROVE YOUR ENVIRONMENT OR ASSESSMENT ALONG THIS DIMENSION? IF SO, PLEASE PROVIDE A BRIEF ACCOUNT OF HOW YOU MIGHT DO THAT, OR HOW THE IDEAS IN THIS WORKING PAPER MIGHT HELP.

(*Use as much space as you need*)

Table 5.3 Ratings of environments and assessments

	ATC21S (N=7)				Grad students (N=11)			
Twenty-first century skills	Mean	SD	Max	Min	Mean	SD	Max	Min
Creativity	7.57	1.81	10	4	5.73	2.53	9	2
Communication	8.00	1.29	9	6	5.50	3.46	9	1
Collaboration	7.86	1.35	9	5	5.59	3.23	9	1
Information literacy	7.57	2.15	9	4	5.55	2.50	10	2
Critical thinking	7.14	1.86	9	4	6.27	3.07	10	2
Citizenship	7.14	2.91	9	2	4.50	2.52	8	1
ICT literacy	7.71	2.69	10	2	4.27	3.10	10	1
Life/career skills	7.57	2.51	9	3	5.86	2.79	10	1
Meta-cognition	8.00	2.00	10	4	4.32	1.95	7	1
Responsibility	7.71	2.21	9	4	4.00	2.76	8	1

Results Obtained by Means of Analytic Templates

Table 5.3 provides descriptive statistics of the ratings of environments and assessments selected by (a) Assessment and Teaching of twenty-first Century Skills project (ATC21S) volunteers versus those selected by (b) graduate students.

Figure 5.6 provides a graphical representation of the ratings of environments and assessments selected by (a) Assessment and Teaching of Twenty-First Century Skills (ATC21S) volunteers versus those selected by (b) graduate students, as listed in Table 5.3.

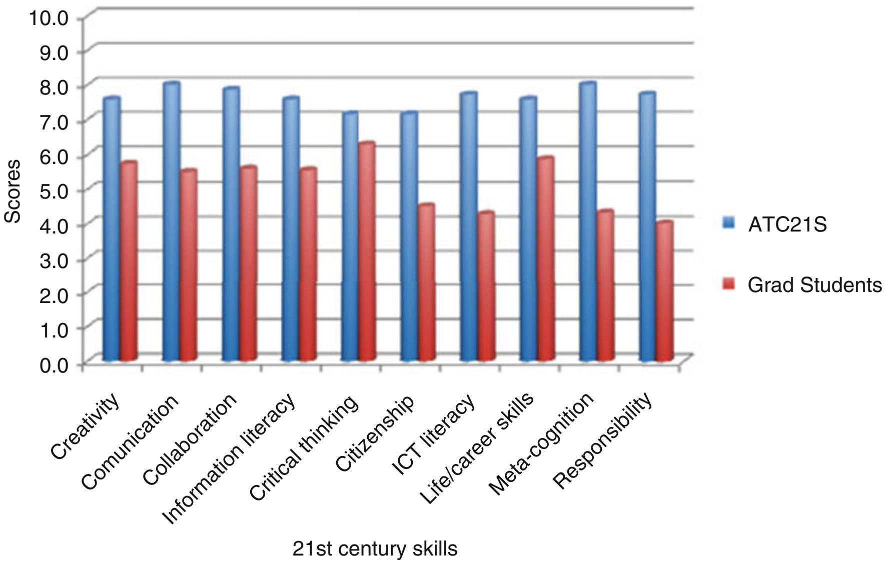


Fig. 5.6 Ratings of environments and assessments

References

Ackoff, R. L. (1974). The systems revolution. *Long Range Planning, 7*, 2–20.

Alexopoulou, E., & Driver, R. (1996). Small group discussion in physics: Peer interaction modes in pairs and fours. *Journal of Research in Science Teaching, 33*(10), 1099–1114.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Anderson, C. (2006). *The long tail: Why the future of business is selling less of more*. New York: Hyperion.

Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership, 57*(5), 13–18.

Arvanitis, S. (2005). Computerization, workplace organization, skilled labour and firm productivity: Evidence for the Swiss business sector. *Economics of Innovation and New Technology, Taylor and Francis Journals, 14*(4), 225–249.

Askenazy, P., Caroli, E., & Marcus, V. (2001). *New organizational practices and working conditions: Evidence from France in the 1990's*. CEPREMAP Working Papers 0106. Downloaded on October 4, 2009, from http://www.cepremap.cnrs.fr/couv_orange/co0106.pdf.

ATC21S – Assessment & Teaching of 21st century skills. (2009). *Transforming education: assessing and teaching 21st century skills* [Assessment Call to Action]. Retrieve from http://atc21s.org/wp-content/uploads/2011/04/Cisco-Intel-Microsoft-Assessment-Call-to-Action.pdf.

Autor, D., Levy, F., & Munane, R. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics, 118*(4), 1279–1334.

Banks, J. A., Au, K. A., Ball, A. F., Bell, P., Gordon, E., Gutierrez, K. D., Brice Heath, S., Lee, C. D., Lee, Y., Mahiri, J., Suad Nasir, N., Valdes, G., & Zhou, M. (2007). *Learning in and out of school in diverse environments: Life-long, life-wide, and life-dee*p. http://www.life-slc.org/

Barron, B. J. (2003). When smart groups fail. *The Journal of the Learning Sciences, 12*(3), 307–35.

Barth, P. (2009). What do we mean by 21st century skills? *American School Board Journal*. Retrieved on October 8, 2009, from http://www.asbj.com/MainMenuCategory/Archive/2009/October/What-Do-We-Mean-by-21st-Century-Skills.aspx

Bateman, H. V., Goldman, S. R., Newbrough, J. R., & Bransford, J. D. (1998). Students' sense of community in constructivist/collaborative learning environments. *Proceedings of the Twentieth Annual Meeting of the Cognitive Science Society* (pp. 126–131). Mahwah: Lawrence Erlbaum.

Bell, D. (1973). *The coming of post-industrial society: A venture in social forecasting*. New York: Basic Books.

Bell, P., Lewenstein, B., Shouse, A. W., & Feder, M. A. (Eds.). (2009). *Learning science in informal environments: People, places, and pursuits*. Washington, DC: National Academies Press.

Bennett, R. E., Persky, H., Weiss, A., & Jenkins, F. (2007). Problem solving in technology rich environments: A report from the NAEP technology-based assessment project, Research and Development Series (NCES 2007–466). U.S. Department of Education, National Center for Educational Statistics. Washington, DC: U.S. Government Printing Office.

Bereiter, C. (1984). How to keep thinking skills from going the way of all frills. *Educational Leadership, 42*(1), 75–77.

Bereiter, C. (2002). *Education and mind in the knowledge age*. Mahwah: Lawrence Erlbaum Associates.

Bereiter, C. (2009). Innovation in the absence of principled knowledge: The case of the Wright Brothers. *Creativity and Innovation Management, 18*(3), 234–241.

Bereiter, C., & Scardamalia, M. (1989). Intentional learning as a goal of instruction. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 361–392). Hillsdale: Lawrence Erlbaum Associates.

Bereiter, C., & Scardamalia, M. (1993). *Surpassing ourselves: An inquiry into the nature and implications of expertise*. Chicago and La Salle: Open Court.

Bereiter, C., & Scardamalia, M. (2006). Education for the knowledge age: Design-centred models of teaching and instruction. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 695–713). Mahwah: Lawrence Erlbaum Associates.

Bereiter, C., & Scardamalia, M. (2009). Teaching how science really works. *Education Canada, 49*(1), 14–17.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., & Rumble, M. (2009). *Developing 21st century skills and assessments*. White Paper from the Assessment and Learning of 21st Century Skills Project.

Black, S. E., & Lynch, L. M. (2003). What's driving the new economy: The benefits of workplace innovation. *The Economic Journal, 114*, 97–116.

Bonk, C. J. (2009). *The world is open: How web technology is revolutionizing education*. San Francisco: Jossey-Bass.

Borghans, L., & ter Weel, B. (2001). *Computers, skills and wages*. Maastricht: MERIT.

Bransford, J. D., & Schwartz, D. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 61–100). Washington, DC: American Educational Research Association.

Bransford, J. D., & Schwartz, D. (2009). It takes expertise to make expertise: Some thoughts about how and why. In K. A. Ericsson (Ed.), *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments* (pp. 432–448). New York: Cambridge University Press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Bransford, J., Mosborg, S., Copland, M. A., Honig, M. A., Nelson, H. G. Gawel, D., Phillips, R. S., & Vye, N. (2009). Adaptive people and adaptive systems: Issues of learning and design. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Eds.), *Second International Handbook of Educational Change. Springer International Handbooks of Education,* (Vol. 23, pp. 825–856). Dordrecht: Springer.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions. *Journal of the Learning Sciences, 2*(2), 141–178.

Brown, A. L., & Campione, J. C. (1996). Psychological theory and design of innovative learning environments: On procedures, principles, and systems. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 289–325). Mahwah: Lawrence Erlbaum Associates.

Carey, S., & Smith, C. (1993). On understanding the nature of scientific knowledge. *Educational Psychologist, 28*(3), 235–251.

Carey, S., Evans, R., Honda, M., Jay, E., & Unger, C. (1989). An experiment is when You Try It and See if It works": A study of junior high school Students' understanding of the construction of scientific knowledge. *International Journal of Science Education, 11*(5), 514–529.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 1*, 33–81.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121–152.

Chuy, M., Scardamalia, M., & Bereiter, C. (2009, August). *Knowledge building and writing development*. Paper presented at the Association for Teacher Education in Europe Conference (ATEE), Palma de Mallorca, Spain.

Collins, A., & Halverson, R. (2009). *Rethinking education in the age of technology: The digital revolution and schooling in America*. New York: Teachers College Press.

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *The Journal of the Learning Sciences, 13*(1), 15–42.

Confrey, J. (1990). A review of research on student conceptions in mathematics, science programming. *Review of Research in Education 16*, 3–55, C.B. Cazden, ed. Washington, DC: American Educational Research Association.

Council, L. (2007). *Skills for the future*. Brussels: Lisbon Council.

Crawford, M. B. (2006). Shop class as soulcraft. *The New Atlantis*, 13, 7–24. Retrieved on October 10, 2009, from http://www.thenewatlantis.com/docLib/20090526_TNA13Crawford2009.pdf.

Crawford, V. M., & Toyama, Y. (2002). *WorldWatcher looking at the environment curriculum: Final external evaluation report*. Menlo Park: SRI International.

Crespi, F., & Pianta, M. (2008). Demand and innovation in productivity growth. *International Review of Applied Economics, 22*(6), 655–672.

Csapó, B. (2007). Research into learning to learn through the assessment of quality and organization of learning outcomes. *The Curriculum Journal, 18*(2), 195–210.

Darling-Hammond, L. (1997). *The right to learn: A blueprint for creating schools that work*. San Francisco: Jossey-Bass.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1).

Darling-Hammond, L., Barron, B., Pearson, P. D., Schoenfeld, A. H., Stage, E. K., Zimmerman, T. D., Cervetti, G. N., & Tilson, J. L. (2008). *Powerful learning: What we know about teaching for understanding*. San Francisco: Jossey-Bass.

David, P. A., & Foray, D. (2003). Economic fundamentals of the knowledge society. *Policy Futures in Education, 1*(1), 20–49.

Dawkins, R. (1996). *The blind watchmaker* (Why the evidence of evolution reveals a universe without design). New York: W. W. Norton.

de Groot, A. D. (1965). *Thought and choice in chess*. New York: Basic Books.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behaviour*. New York: Plenum.

Dickerson, A., & Green, F. (2004). The growth and valuation of generic skills. *Oxford Economic Papers, 56*, 371–406.

Drucker, P. F. (1968). *The age of discontinuity: Guidelines to our changing society*. New York: Harper & Row.

Drucker, P. (1985). *Innovation and entrepreneurship: Practice and principles*. New York: Harper and Row.

Drucker, P. F. (1994, November). The age of social transformation. *Atlantic Monthly,* pp. 53–80.

Drucker, P. F. (2003). *A functioning society: Selection from sixty-five years of writing on community, society, and polity*. New Brunswick: Transaction Publishers.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040–1048.

Earl, L. M. (2003). *Assessment as learning. Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin Press.

Earl, L. M., & Katz, S. (2006). *Leading schools in a data-rich world: Harnessing data for school improvement*. Thousand Oaks: Corwin Press.

Engle, R. A., & Conant, F. R. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and Instruction, 20*(4), 399–483.

Ericsson, K. A. (Ed.). (2009). *Development of professional expertise. Toward measurement of expert performance and design of optimal learning environments*. New York, NY: Cambridge University Press.

Erstad, O. (2008). Trajectories of remixing—Digital literacies, media production and schooling. In C. Lankshear & M. Knobel (Eds.), *Digital literacies. Concepts, policies and practices* (pp. 177–202). New York: Peter Lang.

Fadel, C. (2008, Summer). Deep dives in 21st century curriculum (pp. 3–5). Retrieved on June 10, 2010, from http://mascd.schoolwires.net/1731106417449990/lib/1731106417449990/Summer%202008/June%20Perspectives.Deep%20Dives.2008.pdf.

Fischer, K. W., & Bidell, T. R. (1997). Dynamic development of psychological structures in action and thought. In R. M. Lerner (Ed.) & W. Damon (Series Ed.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (5th ed., pp. 467–561). New York: Wiley.

Frederiksen, J. R., & Collins, A. (1989). A system approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Fujimura, J. (1992). Crafting science: Standardized packages, boundary objects, and translation. In A. Pickering (Ed.), *Science as practice and culture*. Chicago: University of Chicago Press.

Gan, Y. C., Scardamalia, M., Hong, H.-Y., & Zhang, J. (2007). Making thinking visible: Growth in graphical literacy, Grades 3 and 4. In C. Chinn, G. Erkens, & S. Puntambekar (Eds.), *Proceedings of the International Conference on Computer Supported Collaborative Learning 2007* (pp. 206–208). Rutgers, The State University of New Jersey, Newark.

Gaskin, I. W. (2005). *Success with struggling readers: The Benchmark School approach*. New York: Guilford.

Gates, D. (2005). Boeing 787: Parts from around world will be swiftly integrated. *The Seattle Times*, September 11, 2005.

Gera, S., & Gu, W. (2004). The effect of organizational innovation and information technology on firm performance. *International Productivity Monitor, 9*, 37–51.

Gillmore, G. M. (1998, December). Importance of specific skills five and ten years after graduation. OEA Research Report 98–11. Seattle: University of Washington Office of Educational Assessment. Retrieved May 12, 2004, from http://www.washington.edu/oea/9811.htm.

Glaser, R. (1991). Expertise and assessment. In M. Wittrock & E. Baker (Eds.), *Testing and cognition* (pp. 17–30). Englewood Cliffs, NJ: Prentice-Hall.

Gloor, P. A. (2006). *Swarm creativity: Competitive advantage through collaborative innovation networks*. Oxford: Oxford University Press.

Goodwin, C., & Goodwin, M. H. (1996). Seeing as a situated activity: Formulating planes. In Y. Engeström & D. Middleton (Eds.), *Cognition and communication at work* (pp. 61–95). Cambridge: Cambridge University Press.

Greeno, J. G. (1991). Number sense as situated knowing in a conceptual domain. *Journal for Research in Mathematics Education, 22*, 170–218.

Hall, R., & Stevens, R. (1995). Making space: A comparison of mathematical work in school and professional design practices. In S. L. Star (Ed.), *The cultures of computing* (pp. 118–145). London: Basil Blackwell.

Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. Stevenson, J. Azuma, & K. Hakuta (Eds.), *Child development and education in Japan* (pp. 262–272). New York: W. H. Freeman.

Hatano, G., & Osuro, Y. (2003). Commentary: Reconceptualizing school learning using insight from expertise research. *Educational Researcher, 32*, 26–29.

Hearn, G., & Rooney, D. (Eds.). (2008). *Knowledge policy. Challenges for the 21st century*. Northampton: Edward Elgar Publishing, Inc.

Herrenkohl, L. R., & Guerra, M. R. (1998). Participant structures, scientific discourse, and student engagement in fourth grade. *Cognition and Instruction, 16*, 433–475.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *Physics Teacher, 30*, 141–158.

Homer-Dixon, T. (2000). *The ingenuity gap*. New York: Knopf.

Honda, M. (1994). *Linguistic inquiry in the science classroom: "It is science, but it's not like a science problem in a book*."Cambridge: MIT Working Papers in Linguistics.

Johnson, P. (2009). The 21st century skills movement. *Educational Leadership, 67*(1), 11–11.

Katz, S., Earl, L. M., & Jaafar, S. B. (2009). *Building and connecting learning communities: The power of networks for school improvement*. Thousand Oaks: Corwin Press.

Kozma, R. B. (2003). Material and social affordances of multiple representations for science understanding. *Learning Instruction, 13*(2), 205–226.

Kozma, R. B., Chin, E., Russell, J., & Marx, N. (2000). The role of representations and tools in the chemistry laboratory and their implications for chemistry learning. *Journal of the Learning Sciences, 9*(3), 105–144.

Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction, 9*, 285–327.

Laferrière, T. (2001). Collaborative teaching and education reform in a networked world. In M. Moll (Ed.), *But it's only a tool! The politics of technology and education reform* (pp. 65–88). Ottawa: Canadian Teachers Federation and Canadian Centre for Policy Alternative.

Laferrière, T., & Gervais, F. (2008). Communities of practice across learning institutions. In C. Kimble, P. Hildreth, & I. Bourdon (Eds.), *Communities of Practice: Creating Learning Environments for Educators*, Vol. 2 (pp. 179–197). Charlotte: Information Age Publishing Inc.

Lai, M., & Law, N. (2006). Peer Scaffolding of Knowledge Building through Collaboration of Groups with Differential Learning Experiences. *Journal of Educational Computing Research, 35*(2), 121–142.

Lamon, M., Secules, T., Petrosino, A. J., Hackett, R., Bransford, J. D., & Goldman, S. R. (1996). Schools for thought: Overview of the project and lessons learned from one of the sites. In L. Schauble & R. Glaser (Eds.), *Innovation in learning: New environments for education* (pp. 243–288). Hillsdale: Lawrence Erlbaum.

Law, N. (2006). Leveraging technology for educational reform and pedagogical innovation: Policies and practices in Hong Kong and Singapore. *Research and Practice in Technology Education and Learning, 1*(2), 163–170.

Law, N., & Wong, E. (2003). Developmental trajectory in knowledge building: An investigation. In B. Wasson, S. Ludvigsen & U. Hoppe (Eds.), *Designing for change in networked learning environments* (pp.57–66). Dordrecht:: Kluwer Academic Publishers.

Law, N., Lee, Y., & Chow, A. (2002). Practice characteristics that lead to "21st century learning outcomes". *Journal of Computer Assisted Learning, 18*(4), 415–426.

Lee, C. D. (1992). Literacy, cultural diversity, and instruction. *Education and Urban Society, 24*, 279–291.

Lee, E. Y. C., Chan, C. K. K., & van Aalst, J. (2006). Students assessing their own collaborative knowledge building. *International Journal of Computer-Supported Collaborative Learning, 1*, 277–307.

Lehrer, R., Carpenter, S., Schauble, L., & Putz, A. (2000). Designing classrooms that support inquiry. In R. Minstrell & E. Van Zee (Eds.), *Inquiring into inquiry learning and teaching in science* (pp. 80–99). Reston: American Association for the Advancement of Science.

Leiponen, A. (2005). Organization of knowledge and innovation: The case of Finnish business services. *Industry and Innovation, 12*(2), 185–203.

Leonard-Barton, D. (1995). *Wellsprings of knowledge: Building and sustaining the sources of innovation*. Boston: Harvard Business School Press.

Maurin, E., & Thesmar, D. (2004). Changes in the functional structure of firms and the demand for skill. *Journal of Labour Economics, 22*(3), 639–644.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 32*, 13–23.

Messina, R., & Reeve, R. (2006). Knowledge building in elementary science. In K. Leithwood, P. McAdie, N. Bascia, & A. Rodrigue (Eds.), *Teaching for deep understanding: What every educator should know* (pp. 110–115). Thousand Oaks: Corwin Press.

Mestre, J. P. (1994). Cognitive aspects of learning and teaching science. In S. J. Fitzsimmons, & L. C. Kerpelman (Eds.), *Teacher enhancement for elementary and secondary science and mathematics: Status, issues, and problems.* (pp.3–1—3–53). NSF 94–80, Arlington: National Science Foundation.

Minstrell, J. (1989). Teaching science for understanding. In L. Resnick & L. Klopfer (Eds.), *Toward the thinking curriculum: Current cognitive research. 1989 Yearbook of the Association for Supervision and Curriculum Development* (pp. 129–149). Washington, DC: Association for Supervision and Curriculum Development.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centred design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20.

Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G., Hafter, A., Hamel, L., Kennedy, C., Long, K., Morrison, A. L., Murphy, R., Pena, P., Quellmalz, E., Rosenquist, A., Songer, N., Schank, P., Wenk, A., & Wilson, M. (2003). *Design patterns for assessing science inquiry* (PADI Technical Report 1). Menlo Park: SRI International, Center for Technology in Learning.

Moll, L. C. (1986a). *Creating strategic learning environments for students: A community-based approach*. Paper presented at the S.I.G. Language Development Invited Symposium Literacy and Schooling, Annual Meeting of the American Educational Research Association, San Francisco.

Moll, L. C. (1986b). Writing as a communication: Creating strategic learning environments for students. *Theory into Practice, 25*, 102–108.

Moses, R. P. (1994). The struggle for citizenship and math/sciences literacy. *Journal of Mathematical Behaviour, 13*, 107–111.

Moss, J. (2005). Pipes, tubes, and beakers: Teaching rational number. In J. Bransford & S. Donovan (Eds.), *How children learn: History, science and mathematics in the classroom* (pp. 309–350). Washington, DC: National Academies Press.

Moss, J., & Beatty, R. (2006). Knowledge building in mathematics: Supporting collaborative learning in pattern problems. *International Journal of Computer Supported Collaborative Learning, 1*(4), 441–465.

Murphy, M. (2002). *Organizational change and firm performance*. OECD Working Papers. Downloaded on October 3, 2009 from http://puck.sourceoecd.org/vl=18659355/cl=20/nw=1/rpsv/workingpapers/18151965/wp_5lgsjhvj7m41.htm.

National Research Council (2000). *How people learn: Brain, mind, experience, and school*. Expanded version; J. D. Bransford, A. L. Brown, & R. R. Cocking (Eds.). Washington, DC: National Academy Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs: Prentice-Hall.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge creating company: How Japanese companies create the dynamics of innovation*. New York: Oxford University Press.

Norman, D. A. (1993). *Things that make us smart*. Reading: Addison-Wesley Publishing Company.

Nunes, C. A. A., Nunes, M. M. R., & Davis, C. (2003). Assessing the inaccessible: Metacognition and attitudes. *Assessment in Education, 10*(3), 375–388.

Ochs, E., Gonzales, P., & Jacoby, S. (1996). “When I come down I’m in the domain state”: Grammar and graphic representation in the interpretive activity of physicists. In E. Ochs, E. A. Schegloff, & S. Thompson (Eds.), *Interaction and grammar* (pp. 328–369). New York: Cambridge University Press.

Paavola, S., & Hakkarainen, K. (2005). The knowledge creation metaphor—An emergent epistemological approach to learning. *Science and Education, 14*, 535–557.

Panel on Educational Technology of the President's Committee of Advisors on Science and Technology (1997, March). *Report to the President on the use of technology to strengthen K-12 education in the United States*. Retrieved on December 1, 2009, from http://www.ostp.gov/PCAST/k-12ed.html.

Partnership for 21st Century Skills. (2009). Retrieved on October 1, 2009, from http://www.21stcenturyskills.org/

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Pilat, D. (2004). *The economic impact of ICT: A European perspective.* Paper presented at a conference on IT Innovation, Tokyo.

Quellmalz, E. S., & Haertel, G. D. (2008). Assessing new literacies in science and mathematics. In D. J. Leu Jr., J. Coiro, M. Knowbel, & C. Lankshear (Eds.), *Handbook of research on new literacies*. Mahwah: Erlbaum.

Quellmalz, E. S., & Kozma, R. (2003). Designing assessments of learning with technology. *Assessment in Education, 10*(3), 389–407.

Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science, 323*, 75–79.

Raizen, S. A. (1997). Making way for technology education. *Journal of Science Education and Technology, 6*(1), 59–70.

Raizen, S. A., Sellwood, P., Todd, R. D., & Vickers, M. (1995). *Technology education in the classroom: Understanding the designed world*. San Francisco: Jossey-Bass.

Redish, E. F. (1996). *Discipline-specific science education and educational research: The case of physics.* Paper prepared for the Committee on Developments in the Science of Learning, for the Sciences of Science Learning: an Interdisciplinary Discussion.

Reich, R. B. (1991). *The work of nations: Preparing ourselves for 21st century capitalism*. New York: A.A. Knopf.

Robinson, A. G., & Stern, S. (1997). *Corporate creativity. How innovation and improvement actually happen*. San Francisco: Berrett-Koehler Publishers, Inc.

Rotherham, A. J. (2008). *21st-century skills are not a new education trend but could be a fad*. Retrieve October 8, 2009, from http://www.usnews.com/articles/opinion/2008/12/15/21st-century-skills-are-not-a-new-education-trend-but-could-be-a-fad.html

Rotherham, A. J., & Willingham, D. (2009). 21st Century skills: The challenges ahead. *Educational Leadership, 67*(1), 16–21.

Saving the rainforest: REDD or dead? (2009). Retrieved on December 19, 2009, from http://edition.cnn.com/2009/WORLD/europe/12/18/un.redd.program.rainforests/index.html

Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.), *Liberal education in a knowledge society* (pp. 67–98). Chicago: Open Court.

Scardamalia, M., & Bereiter, C. (2003). Knowledge building. In *Encyclopedia of education* (2nd ed., pp. 1370–1373). New York: Macmillan Reference.

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 97–118). New York: Cambridge University Press.

Scardamalia, M., Bereiter, C., Brett, C., Burtis, P. J., Calhoun, C., & Smith Lea, N. (1992). Educational applications of a networked communal database. *Interactive Learning Environments, 2*(1), 45–71.

Schauble, L., Glaser, R., Duschl, R. A., Shulze, S., & John, J. (1995). Students' understanding of the objectives and procedures of experimentation in the science classroom. *Journal of the Learning Sciences, 4*, 131–166.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*(4), 475–522.

Senge, P. M. (1990). *The fifth discipline*. London: Century Business.

Shutt, K., Phillips, R., Van Horne, K., Vye, N., & Bransford, J. B. (2009). *Developing science inquiry skills with challenge-based, student-directed learning*. Seattle: Presentation to the LIFE Center: Learning in Informal and Formal Environments, University of Washington.

Shutt, K., Vye, N., & Bransford, J. D. (2011, April). *The role of agency and authenticity in argumentation during science inquiry*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Orlando, FL.

Simonton, D. K. (1999). *Origins of genius: Darwinian perspectives on creativity*. New York: Oxford University Press.

Smith, C. L., & Wenk, L. (2006). Relations among three aspects of first-year college Students' epistemologies of science. *Journal of Research in Science Teaching, 43*(8), 747–785.

Smith, C. L., Maclin, D., Houghton, C., & Hennessey, M. G. (2000). Sixth-grade Students' epistemologies of science: The impact of school science experiences on epistemological development. *Cognition and Instruction, 18*(3), 349–422.

Spiro, R. J., Vispoel, W. L., Schmitz, J., Samarapungavan, A., & Boeger, A. (1987). Knowledge acquisition for application: Cognitive flexibility and transfer in complex content domains. In B. C. Britton & S. Glynn (Eds.), *Executive control processes in reading* (pp. 177–199). Hillsdale: Lawrence Erlbaum Associates.

Spiro, R. J., Feltovich, P. L., Jackson, M. J., & Coulson, R. L. (1991). Cognitive flexibility, constructivism, and hypertext: Random access instruction for advanced knowledge acquisition in ill-structured domains. *Educational Technology, 31*(5), 24–33.

Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge*. Cambridge: MIT Press.

Stewart, I., & Golubitsky, M. (1992). *Fearful symmetry: Is God a geometer?* Oxford: Blackwell Publishers.

Stipek, D. (2002). *Motivation to learn: Integrating theory and practice* (4th ed.). Needham Heights: Allyn and Bacon.

Stiroh, K. J. (2003). Growth and innovation in the new economy. In D. Jones (Ed.), *New economy handbook* (pp. 723–751). San Diego/London: Elsevier/Academic Press.

Suchman, L. A., & Trigg, R. H. (1993). Artificial intelligence as craftwork. In S. Chaiklin & J. Lave (Eds.), *Understanding practice: Perspectives on activity and context* (pp. 144–178). New York: Cambridge University Press.

Sun, Y., Zhang, J., & Scardamalia, M. (2008). Knowledge building and vocabulary growth over two years, Grades 3 and 4. *Instructional Science*. doi:10.1007/s11251-008-9082-5.

Sun, Y., Zhang, J., & Scardamalia, M. (2010). Developing deep understanding and literacy while addressing a gender-based literacy gap. *Canadian Journal of Learning and Technology 36*(1). Published online at http://www.cjlt.ca/index.php/cjlt/article/view/576

Svihla, V., Vye, N. J., Brown, M., Philips, R., Gawel, D., & Bransford, J. D. (2009). Interactive learning assessments for the 21st century. *Education Canada, 49*(3), 44–47.

Tabak, I., & Baumgartner, E. (2004). The teacher as partner: Exploring participant structures, symmetry, and identity work in scaffolding. *Cognition and Instruction, 22*(4), 393–429.

Teplovs, C. (2008). The knowledge space visualizer: A tool for visualizing online discourse. In G. Kanselaar, V. Jonker, P. A. Kirschner, & F. J. Prins (Eds.), *Proceedings of the International Conference of the Learning Sciences 2008: Cre8 a learning world*. Utrecht: International Society of the Learning.

The North American Council for Online Learning & the Partnership for 21st Century Skills. (2006). *Virtual Schools and 21st Century Skills*. Retrieved on October 8, 2009, from http://www.inacol.org/research/docs/NACOL_21CenturySkills.pdf

Toffler, A. (1990). *Power shift. Knowledge, wealth, and violence at the edge of the 21st century*. New York: Bantam Books.

Trevinarus, J. (1994). Virtual reality technologies and people with disabilities. *Presence: Teleoperators and Virtual Environments, 3*(3), 201–207.

Trevinarus, J. (2002). Making yourself at home—Portable personal access preferences. In K. Miesenberger, J. Klaus, & W. Zagler (Eds.), *Proceedings of the 8th International Conference on Computers Helping People with Special Needs* (pp. 643–648). London: Springer.

Trilling, B., & Fadel, C. (2009). *21st Century skills: Learning for life in our times*. San Francisco: Jossey-Bass.

Tucker, B. (2009). *The Next Generation of Testing*. Retrieved on December 10, 2009, from http://www.ascd.org/publications/educational_leadership/nov09/vol67/num03/The_Next_Generation_of_Testing.aspx.

Tzou, C., & Bell, P. (2010). *Micros and me: Leveraging students' cultural repertoires of practice around microbiology and health in the redesign of a commercially available science kit*. Paper presented at the meeting of the American Educational Research Association, Denver.

U.S. Department of Commerce, U.S. Department of Education, U.S. Department of Labour, National Institute of Literacy, and the Small Business Administration (1999). Report retrieved on October 8, 2009, from http://www.inpathways.net/_ACRNA/21stjobs.pdf

UNESCO. (2005). *Towards knowledge societies*. Paris: United Nations Educational, Scientific, and Cultural Organization.

Venezky, R. L., & Davis, C. (2002). "Quo Vademus? The Transformation of Schooling in a Networked World." Version 8c. OECD Centre for Educational Research and Innovation, Paris. http://www.oecd.org/dataoecd/48/20/2073054.pdf.

Vosniadou, S., & Brewer, W. F. (1989). *The concept of the Earth's shape: A study of conceptual change in childhood*. Unpublished paper. Center for the Study of Reading, University of Illinois, Champaign.

Vygotsky, L. S. (1962). Thought and language. (E. Hanfmann & G. Vakar,Trans.). Cambridge, MA: MIT Press (Original work published in 1934).

Wertime, R. (1979). Students' problems and "courage spans. In J. Lockhead & J. Clements (Eds.), *Cognitive process instruction*. Philadelphia: The Franklin Institute Press.

Wertsch, J. (1998). *Mind as action*. New York: Oxford University Press.

Wiggins, G. P., & McTighe, J. (1997). *Understanding by Design*. Alexandria: Association for Supervision and Curriculum Development.

Wiggins, G. P., & McTighe, J. (2006). Examining the teaching life. *Educational Leadership, 63*, 26–29.

Williams, S. M. (2009). The impact of collaborative, Scaffolded Learning in K-12 Schools: A Meta-Analysis. Report commissioned to The Metiri Group, by Cisco Systems.

Willingham, D. (2008, December 1). Education for the 21st century: Balancing content knowledge with skills. Message posted to http://www.britannica.com/blogs/2008/12/schooling-for-the-21st-century-balancing-content-knowledge-with-skills/

Wilson, B. G. (Ed.). (1996). Constructivist learning environments: Case studies in instructional design. Englewood Cliffs, New Jersey: Educational Technology Publications, Inc.

Wilson, E. O. (1999). *Consilience: The Unity of Knowledge*. London: Vintage Books.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181–208.

Wiske, M. S. (1998). What is teaching for understanding? In M. S. Wiske (Ed.), *Teaching for understanding: Linking research with practice* (pp. 61–86). San Francisco: Jossey-Bass Publishers.

Zhang, J., Scardamalia, M., Lamon, M., Messina, R., & Reeve, R. (2007). Socio-cognitive dynamics of knowledge building in the work of nine- and ten-year-olds. *Educational Technology Research and Development, 55*(2), 117–145.

Zhang, J., Scardamalia, M., Reeve, R., & Messina, R. (2009). Designs for collective cognitive responsibility in knowledge building communities. *The Journal of the Learning Sciences, 18*, 7–44.

Zohgi, C., Mohr, R., & Meyer, P. (2007). *Workplace organization and innovation (Working Paper 405).* Washington, DC: Bureau of Labour Statistics.

Chapter 6
Policy Frameworks for New Assessments

Linda Darling-Hammond

Abstract Many nations around the world have undertaken wide-ranging reforms of curriculum, instruction, and assessment with the intention of better preparing all children for the higher educational demands of life and work in the twenty-first century. While large-scale testing systems in some countries emphasize multiple-choice items that evaluate recall and recognition of discrete facts, there is growing use in many countries of more sophisticated approaches. These approaches include not only more analytical selected response items, but also open-ended items and curriculum-embedded tasks that require students to analyze, apply knowledge, and communicate more extensively, both orally and in writing. A growing emphasis on project-based, inquiry-oriented learning has led to increasing prominence for school-based tasks in state and national systems, taking in research projects, science investigations, use of technology to access information and solve authentic problems, development of products, and presentations about these efforts.

This chapter briefly describes the policy frameworks for assessment systems in Australia, Finland, Singapore, and the UK, with special attention given to identifying cases where assessment of twenty-first century skills has been or may be developed in assessment systems that report information at the national or state, as well as local, levels.

Many nations around the world have undertaken wide-ranging reforms of curriculum, instruction, and assessment, with the intention of better preparing all children for the higher educational demands of life and work in the twenty-first century. To varying degrees, curriculum guidance and assessment systems have begun to focus on a range of twenty-first century skills: the ability to find and organize information

L. Darling-Hammond (✉)
Stanford University, School of Education,
Stanford, California
e-mail: ldh@stanford.edu

P. Griffin et al. (eds.), *Assessment and Teaching of 21st Century Skills*,
DOI 10.1007/978-94-007-2324-5_6, © Springer Science+Business Media B.V. 2012

to solve problems, frame and conduct investigations, analyze and synthesize data, apply learning to new situations, self-monitor, improve one's own learning and performance, communicate well in multiple forms, work in teams, and learn independently.

This interest is also increasingly captured in PISA assessments, which attend explicitly to a number of these twenty-first century skills, going beyond the question posed by many contemporary standardized tests, "Did students learn what we taught them?" to ask, "What can students do with what they have learned?" (Stage 2005). PISA defines literacy in mathematics, science, and reading as students' ability to *apply* what they know to new problems and situations. TIMSS also tests the cognitive domains of applying and reasoning in most items in both 4th grade (60% of items) and 8th grade (65% of items). The IEA's test of reading, PIRLS, focuses on four processes of reading comprehension, with somewhat more weight given to making inferences and integrating ideas and information. This kind of higher-order learning is increasingly emphasized in many nations' assessment systems, in addition to international assessments.

While large-scale testing systems in some countries emphasize multiple-choice items that evaluate recall and recognition of discrete facts, in many countries there is a growing use of more sophisticated approaches, including not only more analytical selected-response items but also open-ended items and curriculum-embedded tasks that require students to analyze, apply knowledge, and communicate more extensively, both orally and in writing. A growing emphasis on project-based, inquiry-oriented learning has led to increasing prominence for school-based tasks in state and national systems, incorporating research projects, science investigations, use of technology to access information and solve authentic problems, development of products, and presentations about these efforts. These assessments, often put together with examination scores, influence the day-to-day work of teaching and learning, focussing it on the development of higher-order skills and the use of knowledge to solve problems.

This paper briefly describes the policy frameworks for assessment systems in four ATC21S countries—Australia, Finland, Singapore, and the UK, with special attention to identifying where assessment of twenty-first century skills has been, or may be, developed in assessment systems that report information at the national or state, as well as at local levels. Identifying the role of twenty-first century skills within these assessment systems serves two purposes. First, this process furthers knowledge about distinct approaches to the integration of twenty-first century skills in countries with different educational governance systems. Second, it provides information about how assessment systems work within the broader policy landscape of each country, which determines student learning opportunities through policies on teacher education and development, as well as on curriculum, instruction, and assessment. With the goal of ensuring that students have the necessary skills to contribute productively to contemporary societies, this chapter offers insights about the ways that different education systems may evolve when supporting an increased focus on twenty-first century skills.

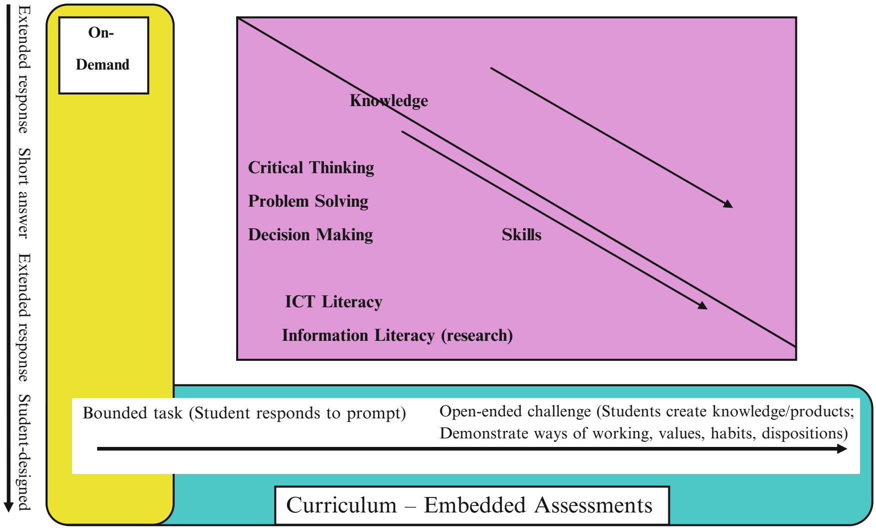

Extended response Short answer Extended response Student-designed
On-Demand
Knowledge
Critical Thinking
Problem Solving
Decision Making
Skills
ICT Literacy
Information Literacy (research)
Bounded task (Student responds to prompt)
Open-ended challenge (Students create knowledge/products; Demonstrate ways of working, values, habits, dispositions)
Curriculum – Embedded Assessments

Fig. 6.1 Contexts for assessing twenty-first century skills

We review the goals and elements of assessment systems in these countries, and how they are implemented in both on-demand tests that occur at relatively brief moments in time as well as classroom-based, curriculum-embedded assessments that may occur over an extended period of time, in which students not only respond to questions or prompts but also construct knowledge products and demonstrate skills through more complex performances. Figure 6.1 seeks to illustrate where, in the context of assessment systems, one might expect to evaluate various kinds of abilities. The list of abilities, presented in Chap. 2, outlines ten kinds of competencies, each of which incorporates dimensions of knowledge, skills, and attitudes or values. The competencies include:

Ways of Thinking

1. Creativity and innovation
2. Critical thinking, problem solving, decision making
3. Learning to learn, metacognition

Ways of Working

4. Communication
5. Collaboration (teamwork)

Tools for Working

6. Information literacy (includes research)
7. ICT literacy

Living in the World

8. Citizenship—local and global
9. Life and career
10. Personal and social responsibility—including cultural awareness and competence

As Fig. 6.1 suggests, certain ways of thinking and uses of tools may be at least partially evaluated with relatively short item on-demand tests, with more extended tasks required for more ambitious forms of problem solving, decision making, and demonstrations of literacy. As one moves from knowledge toward demonstration of skills, as well as attitudes, values, and dispositions—and as one moves closer to examining creativity and innovation, and ways of working and living in the world—the need for more open-ended and extended opportunities to demonstrate abilities becomes more prominent. The most authentic, complex, and applied demonstrations of skills such as unstructured inquiry and problem solving, learning to learn, creativity, communication, collaboration, citizenship, and personal and social responsibility must be examined in contexts that allow larger-scale tasks to be tackled over a longer period of time with more performance-based demonstrations of results than on-demand tests allow. Thus, classroom-based, curriculum-embedded assessments take on an important role in the evaluation of many, perhaps all, of the twenty-first century skills (One could also imagine contexts in which these kinds of assessments would take place in classrooms, as well as in internships or other contexts of employment or life.).

In what follows, we discuss the ways in which assessment systems in four nations provide various kinds of affordances for evaluating twenty-first century skills. In the process, we note that, while smaller countries often have a system of national standards, sometimes accompanied by national tests, larger nations—like Australia, Canada, China, and the USA—have typically had standards and assessment systems at state or province level. In large countries, managing assessment not nationally, but at the state where it remains relatively close to the schools, has often been an important way of managing an integrated system of curriculum, teaching, learning, and assessment. This approach enables strong teacher participation in the assessment process and allows curriculum-embedded assessments to be moderated to ensure consistency in scoring. Smaller nations, which are about the same size as these states or provinces, have been able to support such integrated systems because of their more manageable size.

Currently, governance arrangements are changing in two different directions. On the one hand, both Australia and the USA are attempting to develop national standards and to launch or revise national tests, while also maintaining state assessment systems. On the other hand, school-based assessments—long the norm in countries like Finland and states like Queensland and Victoria in Australia—are becoming increasingly important parts of the assessment systems in jurisdictions like Singapore, England, and Hong Kong, China.

Although this paper does not discuss the new assessment system in Hong Kong, it is perhaps worth noting here that the government's decision to replace the Hong Kong Certificate of Education Examinations with a new Hong Kong Diploma of

Secondary Education places increased emphasis on school-based assessments. As outlined in Hong Kong's "Learning to Learn" reform plan, the goal of the reforms is to shape curriculum and instruction around critical thinking, problem-solving, self-management skills, and collaboration. A particular concern is the development of metacognitive skills, so students may identify their strengths and areas that need additional work (Education Bureau, September 2001; Chan et al. 2008). The Hong Kong Education Examinations Authority explained the rationale for growing use of school-based assessments (SBA) in this way:

> The primary rationale for SBA is to enhance the validity of the assessment, by including the assessment of outcomes that cannot be readily assessed within the context of a one-off public examination…. Obtaining assessments based on student performance over an extended period of time … provides a more reliable assessment of each student.…. Teachers know that SBA, which typically involves students in activities such as making oral presentations, developing a portfolio of work, undertaking fieldwork, carrying out an investigation, doing practical laboratory work or completing a design project, helps students to acquire important skills, knowledge and work habits that cannot readily be assessed or promoted through paper-and-pencil testing. Not only are they outcomes that are essential to learning within the disciplines, they are also outcomes that are valued by tertiary institutions and by employers. Moreover, they are activities that students find meaningful and enjoyable (HKEAA 2009).

In the nations discussed here, school-based assessments often complement centralized "on-demand" tests, constituting 20% to 60% of the final examination score. Tasks are mapped to curriculum expectations or standards and are selected because they represent critical skills, topics, and concepts that cannot be measured in a few hours by an on-demand test. The tasks may be designed and scored locally based on common specifications and evaluation criteria, or they may be designed or scored externally. Whether locally or centrally developed, administration of these tasks occurs at the classroom level, allowing students to engage in intellectually challenging work that taps many of the most ambitious twenty-first century skills, while allowing teachers to obtain immediately available, rich information about the learning process that can inform instruction, something that traditional standardized tests cannot do.

In addition, as teachers use and evaluate these tasks, they can become more knowledgeable about both the standards and how to teach them, and about their students' learning needs. Thus, by improving, the quality of teaching and learning, these forms of assessment may assist in the development of complex abilities in students, as well measuring their abilities. (A summary of assessment system features for the four countries discussed here is shown in Table 6.1, above.)

Australia

Australia is a federation of six states and two territories. The prime responsibility for education is vested in the states and territories under the Australian constitution. In recent years, a more national approach to education has emerged. Currently, state

Table 6.1 International examples of assessment systems

Country/state	Description of core system	What kinds of assessments are used?	Who designs and grades the assessments?
Australia	At the national level, a literacy and numeracy assessment is given at grades 3, 5, 7, and 9. Sample assessments occur in science, ICT literacy, and civics and citizenship. States and localities manage their own assessment systems	*National*—Multiple-choice, short-answer, and extended written responses	*National*—Designed, administered, and scored by the Curriculum Corporation with questions and prompts contributed by state education agencies
Queensland, Australia	All additional assessments are school-based, developed by teachers and based on the national curriculum guidelines and state syllabi On an optional basis, schools may draw on a bank of “rich tasks” from the New Basics project that can be administered across grade levels and scored at the local level, with moderation	*School-based*—Open-ended papers, projects, and inquiries – Rich tasks are complex, interdisciplinary tasks requiring research, writing, and the development of multifaceted products	*School-based*—Assessments are developed, administered, and scored by teachers. Scoring is moderated by regional panels of teachers and professors who examine scored portfolios of student work representing each score point from each grade level from each school. A state panel also looks at specimens across schools as well. Based on these moderation processes, schools are given instructions to adjust grades for comparability – Rich tasks are developed by teachers with assessment developers; they are accompanied by scoring rubrics and moderation processes by which the quality of student work and scoring can be evaluated

Victoria, Australia	All additional assessments are school-based until 11th and 12th grades, when students choose to take exams in different subject areas as part of the Victorian Certificate of Education (VCE), used to provide information to universities and employers. The VCE exams have both external and school-based components. At least 50% of the total examination score is comprised of required classroom-based assignments and assessments given throughout the school year Schools have access to an on-demand assessment system for students in years 3–10 which includes computer adaptive literacy and numeracy tests that score students according to a statewide standards scale All students on entry to school and at end of prep, year 1, and year 2 complete an online assessment of English (The English Online Interview). Also available for on-demand testing by teachers in primary school is the mathematics online interview, providing rich diagnostic information about individual student learning. The mathematics online interview is used optionally by teachers of prep to year 2 students, with an estimated 70% of schools routinely using this assessment for prep students	*State VCE*—Multiple-choice (25%) and open-ended (75%) written, oral, and performance elements *School-based*—Lab experiments, essay, research papers and presentations *On-entry, prep–year 2*—Oral language, phonemic awareness, fluency, reading, comprehension, writing, spelling – Mathematics online interview	The Victoria Curriculum and Assessment Authority (VCAA) establishes courses in a wide range of studies, oversees the development of the external examinations by teachers and university faculty, and ensures the quality of the school-assessed component of the VCE. Teachers score the open-ended items on the external exam and design and score the classroom-based assessments in response to syllabus guidelines. Online marking has been introduced for one examination and will be used for more examinations in the future. Online marking has been introduced due to efficiencies it provides as well as enhanced quality control of marking. The quality of the tasks assigned by teachers, the work done by students, and the appropriateness of the grades and feedback given to students are audited through an inspection system; schools are given feedback on all of these elements. In addition, the VCAA uses statistical moderation based on the external exam scores to ensure that the same assessment standards are applied to students across schools The prep to year 2 English online interview has been designed specifically to provide an indication of student achievement against the Victorian Essential Learning Standards (VELS).). It is administered and marked by classroom teachers via an Internet-based system

(continued)

Table 6.1 (continued)

Country/state	Description of core system	What kinds of assessments are used?	Who designs and grades the assessments?
Finland	Student performance is evaluated on a sample basis by the Finnish education authorities at the end of 2nd and 9th grades to inform curriculum and school investments All other assessments are designed and managed locally based on the national curriculum A voluntary matriculation examination is taken by most students to provide information to colleges. Students choose which subjects they will sit for (usually at least four), with the test in the students' mother tongue being compulsory	*National*—Problems and written tasks that ask students to apply their thinking *School-based*—Papers, research tasks, & presentations The tests use mostly open-ended questions to evaluate skills including problem-solving, analysis, and writing	*National*—Designed by teachers through the Finnish Ministry of Education. Graded by teachers *School-based*—Teachers design and grade tasks based on recommended assessment criteria and benchmarks for each subject and grade within the national core curriculum The exam is administered, organized, and evaluated by the Matriculation Exam Board appointed by the Finnish Ministry of Education. Teachers grade the matriculation exams locally by using the official guidelines, and samples of the grades are reexamined by professional raters hired by the exam board
Singapore	External examinations are given at the end of primary school (grade 6) in mathematics, science, English, and mother tongue (Malay, Chinese, or Tamil). Results are used to guide course placements in secondary school All other assessments are school-based After 4 years of secondary school, students take the GCE N- or O-level examinations. Students choose the elective subject areas in which they want to be examined. Exams have school-based components that comprise up to 20% of the final score. Results are used as information for postsecondary education. GCE A-level examinations may be taken after 2 years of tertiary education	*National*—Short and long open-ended responses *School-based*—Coursework, research projects, investigations *National*—Short and long open-ended responses and multiple-choice items *School-based*—Research projects, laboratory investigations	*National*—The Singapore Education Assessment Board designs the assessments and manages the assessment system *School-based*—Designed and graded by the classroom teacher in response to the syllabus *National*—The Singapore Education Assessment Board manages the assessment system. The GCE examinations are developed by the Cambridge International Examinations Group *School-based*—Teachers develop, implement, and score projects and other products that complement the external examinations

United Kingdom	National curriculum assessments are enacted primarily as guidance for school-based formative and progress assessments conducted by teachers. A mandatory set of assessments at ages 7 and 11 includes externally developed tasks and observation scales implemented by teachers. Teachers choose which tasks and tests to use and when to use them, within certain parameters Assessments for primary school are designed and managed locally based on the national curriculum and guidance provided through the Assessing Pupil Progress (APP) program	*National*—Observation scales completed by teachers regarding pupils' work and performance on specific kinds of tasks; written, oral, and performance tasks and tests *School-based*—Coursework, tests, projects, essays	*National*—The Qualifications and Curriculum Authority (QCA) manages and develops the national assessments which are implemented and scored by teachers. They also provide a range of guidance and support for in-school assessment *School-based*—Teachers evaluate student performance and work samples based on the national curriculum and syllabi. Extensive guidance for documenting pupil performance and progress, with indicators showing relationships to national standards are provided through the Assessing Pupils' Progress project. Regional authorities support teacher training for assessment and in-school moderation
	Most students take a set of exams at year 11 (age 16) to achieve their General Certificate of Secondary Education (GCSE). If they take advanced courses, they may later take A-level exams, which provide information to universities. Students choose the exams they will take based on their interests and areas of expertise. About 40–75% of the exam grade is based on externally developed tests and 25–60% is school-based	*National*—Essays and open-ended problem solutions, oral language assessments *School-based*—Coursework, tests, projects	*National*—External exams are designed and graded by examining groups serving different schools (e.g., Oxford Cambridge, Ed Excel, the Assessments and Qualifications Alliance) *School-based*—Teachers develop and score school-based components based on the syllabus

and territory governments are responsible for developing policy, delivering services, monitoring and reviewing performance of individual schools, and regulating schools so as to work toward national objectives and achievement of outcomes compatible with local circumstances and priorities. The Australian Government provides support for schooling through general recurrent, capital and targeted programs, policy development, research and analysis of nationally significant education issues. A key priority for the government is to provide leadership toward achieving a nationally consistent school system through common national testing in key subject areas and consistency in curriculum outcomes. While state and territory governments provide the majority of recurrent funding to government schools, the Australian Government is the primary funding source of the non-government schooling sector.

At the national level, in recognition that students need to be prepared for the higher educational demands of life and work in the twenty-first century, the Australian Government, in partnership with state and territory governments, has embarked upon a series of national reforms in education. Key aspects of these reforms that are relevant to AT21CS are outlined below:

National Efforts

Assessment

The establishment of the Australian Curriculum, Assessment and Reporting Authority (ACARA) brings together the management of curriculum, assessment, and reporting for the first time at the national level. This is intended to help streamline and simplify national education governance, which in turn is expected to help reduce duplication of resources and costs and provide a central mechanism through which Australian government can drive national priorities in education.

A new National Assessment Program (NAP), managed by ACARA, includes annual national literacy and numeracy assessments and triennial national sample assessments in science literacy, civics and citizenship, and ICT literacy. Australia's participation in international assessments (PISA, TIMSS, and PIRLS) is also included in this suite of NAP assessments, but is managed separately. As part of its 2010 work program, ACARA will be undertaking a review of the NAP sample assessments, which may present an opportunity to incorporate AT21CS project outcomes.

The reading, language conventions, and numeracy NAP tests consist mostly of multiple-choice items (about 75% of items), with some short constructed responses where relevant. The writing test is a longer constructed response where students are required to write on a specified topic and genre. The NAP sample assessments, which are administered to a representative sample of students from each state and territory across school sectors, include tests in science literacy (NAPSL) at year 6, civics and citizenship (NAPCC) at years 6 and 10, and ICT Literacy (NAPICTL) at years 6 and 10. These assessments are conducted on a rolling triennial basis.

A selection of items from the sample tests, those not required for equating purposes, are available for schools that wish to use them to assess their students if they wish.

In addition to multiple choice and short answer items, the science literacy test includes a group practical task. Information from the group practical task is used by individual students to answer items; the practical task itself is not marked and collaboration is not specifically assessed. The ICT literacy test requires students to use computers, mostly online, as part of the assessment process. Students are required to put together pieces of work using simulated web information and specific computer programs such as a word processor, spreadsheet, and presentation program.

The Australian Government is currently undertaking a project to evaluate the usefulness of existing information and communications (ICT)-based assessment tools and resources for national curriculum key learning areas. In addition, it is proposed that the research will document ICT-based assessment tools and resources in the vocational education and training (VET) and higher education resources, as well as similar tools and resources from selected overseas countries. This research will provide vital information to assist the Australian Government to maximize the opportunities to enrich teaching and learning with the use of ICT tools and resources. It is expected that this project will be informed by the work being undertaken as part of the AT21CS.

Curriculum

The current education landscape across Australia is varied and complex; each state and territory has its own curriculum in place, assessment and reporting arrangements that have been built over time and in response to local considerations. The national curriculum being developed by ACARA seeks to equip young Australians with the skills, knowledge, and capabilities they need to engage with and prosper in society, compete in a globalized world, and thrive in the information-rich workplaces of the future.

ACARA recognizes that not all learning is limited to the learning areas into which the school curriculum has traditionally been divided.[1] Accordingly, the national curriculum includes ten general capabilities to be addressed across the curriculum, which aim to develop twenty-first century skills. These are: literacy, numeracy, information and communication technology (ICT), thinking skills, creativity, self-management, teamwork, intercultural understanding, ethical behavior, and social competence.

[1] See ACARA, *The Shape of the Australian Curriculum*. Available http://www.acara.edu.au/publications.html

Teaching

The Smarter Schools Improving Teacher Quality National Partnership (TQNP)[2] provides funding for reforms to attract, train, place, develop, and retain quality teachers and school leaders. These reforms include implementing a standards-based National Teaching Professional Framework that will provide nationally consistent requirements and principles for accrediting teachers at the graduate, competent, highly accomplished, and leading teacher levels, as well as enhancing professional learning and performance appraisal for teachers and school leaders throughout their careers. This framework will also support nationally consistent teacher registration and improvements in the quality of teacher training, by accrediting pre-service education courses. Other components of the TQNP include professional development and support initiatives to empower principals to be better able to manage their schools to meet the needs of their students, mechanisms to attract high-quality graduates to teaching, and measures to improve teacher retention by rewarding quality teachers and school leaders and improving the quality of teacher workforce data.

In addition to the framework, the Australian State and Territory Education Ministers have agreed to establish the Australian Institute for Teaching and School Leadership (AITSL). AITSL will promote excellence in the profession of teaching and school leadership by:

- Developing and overseeing a set of national standards for teaching and school leadership and implementing an agreed system of national accreditation of teachers based on these standards; and
- Promoting excellence and national leadership in the professional development of teachers and school leaders.

A priority of AITSL is to advise on the delivery of world-leading professional development and provide support for it, empowering principals to better manage their schools to achieve improved student results.

Technology

Through a major Digital Education Revolution (DER) initiative, the Australian Government is providing $2.2 billion over 6 years to:

- Provide for new information and communication technology (ICT) equipment for all secondary schools with students in years 9–12, through the National Secondary School Computer Fund;
- Support the deployment of high-speed broadband connections to Australian schools;

[2] Further information at www.deewr.gov.au/Schooling/Programs/SmarterSchools/Pages/default.aspx

- Collaborate with states and territories and Deans of Education to ensure that new and continuing teachers have access to training in the use of ICT that enables them to enrich student learning;
- Provide for online curriculum tools and resources that support the national curriculum and specialist subjects such as languages;
- Enable parents to participate in their child's education through online learning and access;
- Set up support mechanisms to provide vital assistance for schools in the deployment of ICT.

State Assessment Systems

In many states, school-based performance assessments targeting many of the twenty-first century skills have been a longstanding part of the system. In some cases, states have also developed centralized assessments with performance components. Here we describe these approaches across states; then we shall look more deeply at exemplars of assessment tasks in two states: Queensland and Victoria. One of these states, Queensland, has a highly developed system of centrally moderated local performance assessments, and the other, Victoria, uses a blended model of centralized and school-based assessments, both of which use moderated scoring.

A number of states have developed assessment systems that provide opportunities for students to demonstrate approaches to problem-solving and the construction of ideas and products. There are also some innovative approaches to supporting the development of productive attitudes, values, and dispositions toward inquiry and innovation, as well as the quality of teaching.

For example, the New South Wales Essential Secondary Science Assessment (ESSA) program (conducted at year 8) is a diagnostic test that contains several extended response tasks, along with multiple-choice items. It also contains an unscored "survey" to assess students' values and attitudes related to science and science learning (A teacher survey and a parent survey are also conducted each year as an addition to the assessment program.). Another aspect of the test that is fully developed but not yet mandatory is an online practical component that simulates a science investigation. Students complete multiple-choice and short-response items, and an extended response task as they conduct their online investigation. It is expected that the pencil-and-paper format of the test will be replaced by a completely online test in 2011.

Teachers mark the three extended response tasks, including that from the online practical component, at marking centers. Results are reported to schools through the NSW DET School Measurement, Assessment and Reporting Toolkit (SMART), a powerful computer package that displays results flexibly and enables the manipulation of data by schools. Curriculum support materials related to test items are available online for participating schools.

In *Western Australia*, external assessments of Science and Society and the Environment occur at grades 5, 7, and 9. In addition, the Curriculum Council establishes courses and examinations in years 11 and 12 across a wide range of disciplines and ensures the quality of the school-assessed component of the Western Australian Certificate of Education (WACE) (similar systems are used in South Australia and Victoria.). External examinations are combined with school-based assessments that range from laboratory experiments, essays, research papers, presentations, demonstrations, and projects to school-based tests and examinations. State external assessments are mainly written examinations, with some courses also having external practical examinations (e.g., oral for languages, instrumental solos for music, visual diaries for visual art, and flight simulation for aviation). In years 11 and 12, the Curriculum Council uses statistical moderation based on the external exam scores to ensure that the same assessment standards are applied to students across schools.

In addition to a syllabus for teachers to use as a reference in developing teaching and learning programs, the Western Australian Department of Education also provides grade standards and student work exemplars to support teachers in making appropriate and consistent judgments about student achievement, along with diagnostic assessment and reporting tools.

Extensive databases are used for administering tests and recording data. Online interactive programs based on the test data facilitate diagnostic assessment, moderation and evaluation of student, cohort, school, and system performance. Substantial scanning technology is used for population-based assessments, including full scanning of writing scripts with sophisticated on-screen marking.

Similarly, in *the Australian Capital Territory (ACT)*, where school-based assessment is the primary approach until grade 10, individual teachers design and grade tasks based on school-developed assessment criteria and curriculum documents. They are guided by the stages of development outlined in the ACT curriculum framework. Students also assess themselves against specific criteria. The assessment of students' use of ICT is embedded across all curriculum areas. It is also an integral part of administration, scoring, moderation, sharing of assessments and student work. A "Myclasses" online resource is available for the sharing of assessment tasks among teachers.

In *South Australia*, interesting progress is being made in creating more comparable evaluation of school-based assessments. Through grade 10, all students are assessed through school assessments developed by teachers, and judgments are made against the outcomes in the South Australian Curriculum, Standards and Accountability (SACSA) Framework. Schools can enter the outcomes data into the SACSA Achievement System software. Curriculum Services of the Department of Education and Children's Services manage a Peer Review Moderation project to promote consistency across schools and to provide quality assurance for the data entered into the SAS, by way of a random sample of schools across subject areas. This project also plans to expand assessment of the SACSA Essential Learnings (identity, interdependence, thinking, futures, and communication). Many schools have assessment programs that incorporate communication, collaboration, critical thinking, citizenship, ICT literacy, and learning to learn.

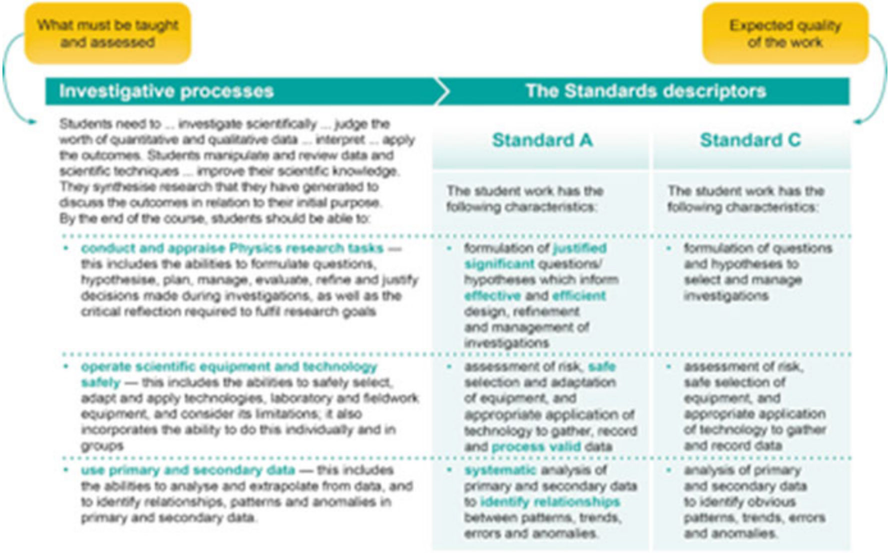
At grades 11–12, a variety of assessment instruments are used in school-based assessments of the South Australia Certificate of Education (SACE). All stage 1 subjects are assessed using wholly school-based assessment. External assessment components, including written examinations, performance and practical examinations, studies, investigations, and oral examinations apply to some stage 2 subjects. When the new SACE is introduced at stage 2 in 2011, all subjects will have 70% school-based and 30% external assessment components. It is intended that a student who completes the SACE will:

- Be an active, confident participant in the learning process (confidence);
- Take responsibility for his or her own learning and training;
- Respond to challenging learning opportunities, pursue excellence, and achieve in a diverse range of learning and training situations;
- Work and learn individually and with others in and beyond school to achieve personal or team goals (independence, collaboration, identity);
- Apply logical, critical, and innovative thinking to a range of problems and ideas (thinking, enterprise, problem-solving, future);
- Use language effectively to engage with the cultural and intellectual ideas of others (communication, literacy);
- Select, integrate, and apply numerical and spatial concepts and techniques;
- Be a competent, creative, and critical user of information and communication technologies (information technology);
- Have the skills and capabilities required for effective local and global citizenship, including a concern for others (citizenship, interdependence, responsibility toward the environment, responsibility toward others);
- Have positive attitudes toward further education and training, employment, and lifelong learning (lifelong learning).

With the introduction of the new SACE, five capabilities (communication, citizenship, personal development, work, and learning) are embedded in all subjects, with some or all of the capabilities being explicitly assessed. The introduction of the new SACE will also offer new opportunities for using technology, including e-Portfolios, e-Assessment, and e-Moderation in addition to an enhanced management system.

Queensland. In Queensland, school-based assessment has been the norm for 40 years. Until the early 1970s, a centralized examination system controlled the curriculum; after it was eliminated, all assessments became school-based. These assessments are developed, administered, and scored by teachers in compliance with the national curriculum guidelines and state syllabi (also developed by teachers), and are moderated by panels that include teachers from other schools and professors from the tertiary education system. Recently, centrally developed tasks and a 12th grade test have been added.

To create the standards used throughout the state, the central authority gathers groups of teachers and subject experts to write standards that specify different levels of achievement and describe the characteristics of student work at each level. In the excerpt from Queensland's science standards shown in Fig. 6.2 below, the left



What must be taught and assessed

Expected quality of the work

Investigative processes	The Standards descriptors	
	Standard A	Standard C
Students need to ... investigate scientifically ... judge the worth of quantitative and qualitative data ... interpret ... apply the outcomes. Students manipulate and review data and scientific techniques ... improve their scientific knowledge. They synthesise research that they have generated to discuss the outcomes in relation to their initial purpose. By the end of the course, students should be able to:	The student work has the following characteristics:	The student work has the following characteristics:
• **conduct and appraise Physics research tasks** — this includes the abilities to formulate questions, hypothesise, plan, manage, evaluate, refine and justify decisions made during investigations, as well as the critical reflection required to fulfil research goals	• formulation of **justified significant** questions/ hypotheses which inform **effective** and **efficient** design, refinement and management of investigations	• formulation of questions and hypotheses to select and manage investigations
• **operate scientific equipment and technology safely** — this includes the abilities to safely select, adapt and apply technologies, laboratory and fieldwork equipment, and consider its limitations; it also incorporates the ability to do this individually and in groups	• assessment of risk, **safe** selection and adaptation of equipment, and appropriate application of technology to gather, record and **process valid** data	• assessment of risk, safe selection of equipment, and appropriate application of technology to gather and record data
• **use primary and secondary data** — this includes the abilities to analyse and extrapolate from data, and to identify relationships, patterns and anomalies in primary and secondary data.	• **systematic** analysis of primary and secondary data to **identify relationships** between patterns, trends, errors and anomalies.	• analysis of primary and secondary data to identify obvious patterns, trends, errors and anomalies.

Fig. 6.2 Excerpt from Queensland science standards

column describes the objectives or "Essential Learnings" that must be taught and assessed by teachers. The objectives convey the knowledge or skill expected at each standard. The standard descriptors to the right detail the expected characteristics and quality of the work. The teachers and experts also develop samples of work as exemplars of the different levels. These standards guide the assessments that teachers develop and their scoring.

The syllabi seek to strike a balance between "informed prescription" and "informed professionalism." They spell out a small number of key concepts and skills to be learned in each course and the kinds of projects or activities (including minimum assessment requirements) students should be engaged in. Each school designs its program to fit the needs and experience of its students, choosing specific texts and topics with this in mind. However, all schools evaluate student work using shared criteria based on the course objectives and specific standards for an A, B, C, D, or E mark.

As the criteria from the physics syllabus in Fig. 6.3 indicate, in the category of *Knowledge and conceptual understanding*, work that meets an "A" standard demonstrates interpretation, comparison, and explanation of complex concepts, theories, and principles, whereas work at an "E" standard is characterized by reproduction of isolated facts and application of simple, given algorithms. In this particular course, objectives also include *Investigative Processes*, and *Evaluating and Concluding,* with indicators spelled out for each. The expectations of work quality are challenging and include critical thinking, problem-solving, decision making, research, and communication skills, as shown in the example in this figure.

In Queensland science courses, students must complete an extended experimental investigation. The instructions for the task read:

Within this category, instruments are developed to investigate a hypothesis or to answer a practical research question. The focus is on planning the extended experimental investigation, problem solving and analysis of primary data generated through experimentation by the student. Experiments may be laboratory or field based. An extended experimental investigation may last from four weeks to the entirety of the unit of work. The outcome of an extended experimental investigation is a written scientific report. *Aspects of each of the three criteria should be evident in the investigation.* For monitoring, the discussion/conclusions/evaluation/recommendations of the report should be between 1500 and 2000 words.

To complete such an investigation the student must:

- develop a planned course of action
- clearly articulate the hypothesis or research question, providing a statement of purpose for the investigation
- provide descriptions of the experiment
- show evidence of modification or student design
- provide evidence of primary and secondary data collection and selection
- execute the experiment(s)
- analyze data
- discuss the outcomes of the experiment
- evaluate and justify conclusion(s)
- present relevant information in a scientific report.

Fig. 6.3 Science assessment, Queensland, Australia

An example from a year 12 paper shows how a student investigated a problem entitled, “The Air Pocket.” The assessment starts with a picture, shown in Fig. 6.4 below, of a vertical air jet from a straw producing a cavity on a water surface.

The student investigated the parameters that would affect the volume of the cavity, preparing a 32-page paper that met the criteria described earlier, including evaluating the problem theoretically and empirically, presenting data through tables and charts, analyzing findings both by summarizing individual results and developing a regression to evaluate the combined effects of several variables on the volume of the cavity, and by evaluating the results, also listing the potential errors and additional research needed. Overall, the paper more closely resembles a research report from a scientific laboratory than a traditional high school physics test. The student concluded:

> It was determined through initial theoretical research that the predominant influences on the cavity’s volume were air speed, diameter of nozzle/straw and distance between straw/nozzle and water. Upon testing the effects of changing an individual parameter with respect to volume, every possible variation was tried, such that eventually a complete set of values was obtained. To combine the different parameters into a single equation, a multiple regression was used; to determine both the constant factor and the powers to which each of the variables should be raised. The resultant r^2 value was 0.96 indicating an excellent fit for the data while the average percentage error was 1.59% and the median percentage error, 6.71%. … [In future experiments], it would be suggested to do the experiments on a larger scale as

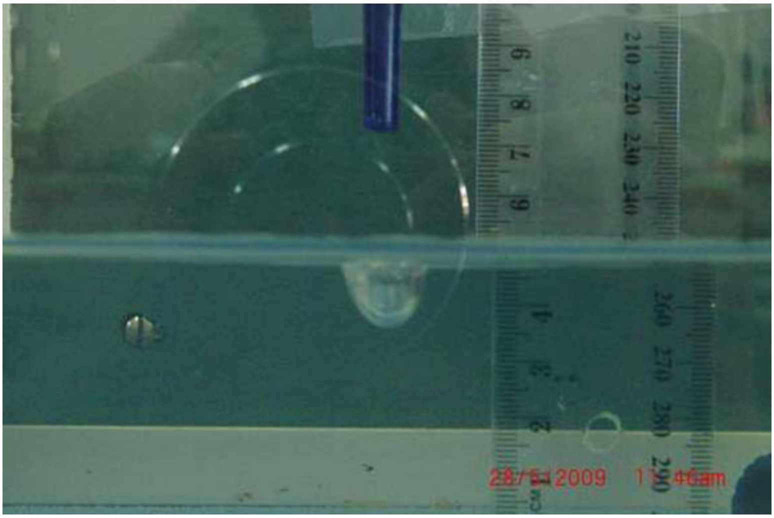


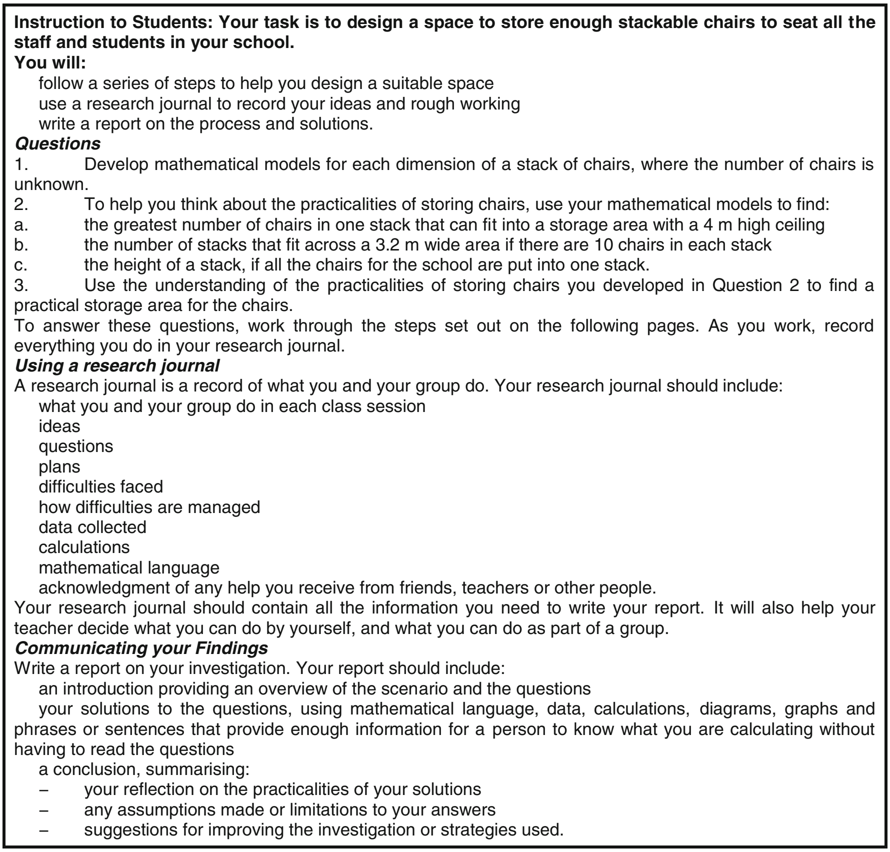
Fig. 6.4 Picture for problem on an air pocket

> this would virtually eliminate the effects of surface tension while cutting down unfounded accuracy in the model (the volume could be measured in cubic centimetres or cubic metres, resulting in a more realistic fit, with data that is not required to be impossibly precise. Finally, it would be suggested to trial the effects of the different orientation of the straw/nozzle, as tilting it would give a completely differently shaped cavity (due to the dispersion characteristics of air).

Thus, students go beyond their own empirical data and conclusions to reflect upon the accuracy of their findings and the means for improving their investigation. These kinds of extended responses are demanded in all of the subject areas, shaped by the core concepts and modes of inquiry of the disciplines. Student reflection is also a common element of the assessments. Consistent scoring of such intellectually ambitious work is made possible by internal and external moderation processes, and by the clear guidance of the syllabi and rubrics used to set standards for the work.

At lower grade levels, the Queensland Studies Authority (QSA) has recently developed and piloted centrally devised Queensland Comparable Assessment Tasks (QCATs) for years 4, 6, and 9 in the English, Mathematics, and Science *Essential Learnings* and *Standards*. These tasks, available in an Assessment Bank, aim to provide authentic, performance-based assessments that can be used to evaluate learning and are scored in moderated processes by teachers to develop comparability of reported results. The task, shown in Fig. 6.5, for grade 9 mathematics, illustrates the kind of problem-solving, critical thinking, collaboration, creativity, and communication evaluated by the tasks.

All of the 98,000 students in Queensland's 11th and 12th grades complete multiple assessments like these, based on the national standards, the state syllabi, and the



Instruction to Students: Your task is to design a space to store enough stackable chairs to seat all the staff and students in your school.

You will:

follow a series of steps to help you design a suitable space
use a research journal to record your ideas and rough working
write a report on the process and solutions.

Questions

1. Develop mathematical models for each dimension of a stack of chairs, where the number of chairs is unknown.
2. To help you think about the practicalities of storing chairs, use your mathematical models to find:
a. the greatest number of chairs in one stack that can fit into a storage area with a 4 m high ceiling
b. the number of stacks that fit across a 3.2 m wide area if there are 10 chairs in each stack
c. the height of a stack, if all the chairs for the school are put into one stack.
3. Use the understanding of the practicalities of storing chairs you developed in Question 2 to find a practical storage area for the chairs.

To answer these questions, work through the steps set out on the following pages. As you work, record everything you do in your research journal.

Using a research journal

A research journal is a record of what you and your group do. Your research journal should include:

what you and your group do in each class session
ideas
questions
plans
difficulties faced
how difficulties are managed
data collected
calculations
mathematical language
acknowledgment of any help you receive from friends, teachers or other people.

Your research journal should contain all the information you need to write your report. It will also help your teacher decide what you can do by yourself, and what you can do as part of a group.

Communicating your Findings

Write a report on your investigation. Your report should include:

an introduction providing an overview of the scenario and the questions
your solutions to the questions, using mathematical language, data, calculations, diagrams, graphs and phrases or sentences that provide enough information for a person to know what you are calculating without having to read the questions
a conclusion, summarising:
– your reflection on the practicalities of your solutions
– any assumptions made or limitations to your answers
– suggestions for improving the investigation or strategies used.

Fig. 6.5 Queensland mathematics assessment: "Stackable chairs"

school's approved work plan. At the end of the year, teachers collect a portfolio of each student's work, which includes the specific assessment tasks, and grade it on a 5-point grading scale. To calibrate these grades, teachers put together a selection of portfolios from each grade level—one from each of the 5 score levels plus borderline cases—and send these to a regional panel for moderation. A panel of five teachers rescores the portfolios and confers about whether the grade is warranted, making a judgment on the spread. State review panels also look at a sample of student work from each district to ensure that schools implement the standards across all districts. Based on this analysis, and on a standardized statewide test called the Queensland Core Skill (QCS) Test, at year 12, the Queensland authority confirms the levels of achievement proposed by school programs and may adjust them if they do not fit the standards.

Aiming for even more applied, interdisciplinary work, Queensland developed a "rich tasks" approach to standards and assessment, which was introduced as a pilot

Students must identify, explore and make judgments on a biotechnological process to which there are ethical dimensions. Students identify scientific techniques used as well as significant recent contributions to the field. They will also research frameworks of ethical principles for coming to terms with an identified ethical issue or question. Using this information they prepare pre-conference materials for an international conference that will feature selected speakers who are leading lights in their respective fields.

In order to do this students must choose and explore an area of biotechnology where there are ethical issues under consideration and undertake laboratory activities that help them understand some of the laboratory practices. This enables them to:

Provide a written explanation of the fundamental technological differences in some of the techniques used, or of potential use, in this area (included in the pre-conference package for delegates who are not necessarily experts in this area).

Consider the range of ethical issues raised in regard to this area's purposes and actions, and scientific techniques and principles and present a deep analysis of an ethical issue about which there is a debate in terms of an ethical framework.

Select six real-life people who have made relevant contributions to this area and write a 150-200 word précis about each one indicating his/her contribution, as well as a letter of invitation to one of them.

This assessment measures research and analytic skills; laboratory practices; understanding biological and chemical structures and systems, nomenclature and notations; organizing, arranging, sifting through, and making sense of ideas; communicating using formal correspondence; précis writing with a purpose; understanding ethical issues and principles; time management, and much more.

Fig. 6.6 A rich task: "Science and ethics confer", Queensland, Australia

in 2003. Part of the "New Basics" project, this effort has created extended, multidisciplinary tasks that are developed centrally and used locally when teachers determine the time is right and they can be integrated with locally oriented curriculum (Queensland Government 2001). These are "specific activities that students undertake that have real-world value and use, and through which students are able to display their grasp and use of important ideas and skills." Rich tasks are defined as

> A culminating performance or demonstration or product that is purposeful and models a life role. It presents substantive, real problems to solve and engages learners in forms of pragmatic social action that have real value in the world. The problems require identification, analysis and resolution, and require students to analyze, theorize and engage intellectually with the world. As well as having this connectedness to the world beyond the classroom, the tasks are also rich in their application: they represent an educational outcome of demonstrable and substantial intellectual and educational value. And, to be truly rich, a task must be transdisciplinary. Transdisciplinary learnings draw upon practices and skills across disciplines while retaining the integrity of each individual discipline.

One task description is summarized in Fig. 6.6 above. A bank of these tasks now exists across grade levels, along with scoring rubrics and moderation processes by which the quality of the tasks, the student work, and the scoring can be evaluated. Studies have found stronger student engagement in schools using the rich tasks. On traditional tests, the "New Basics" students scored about the same as students in the traditional program, and they scored notably better on assessments designed to gauge higher order thinking.

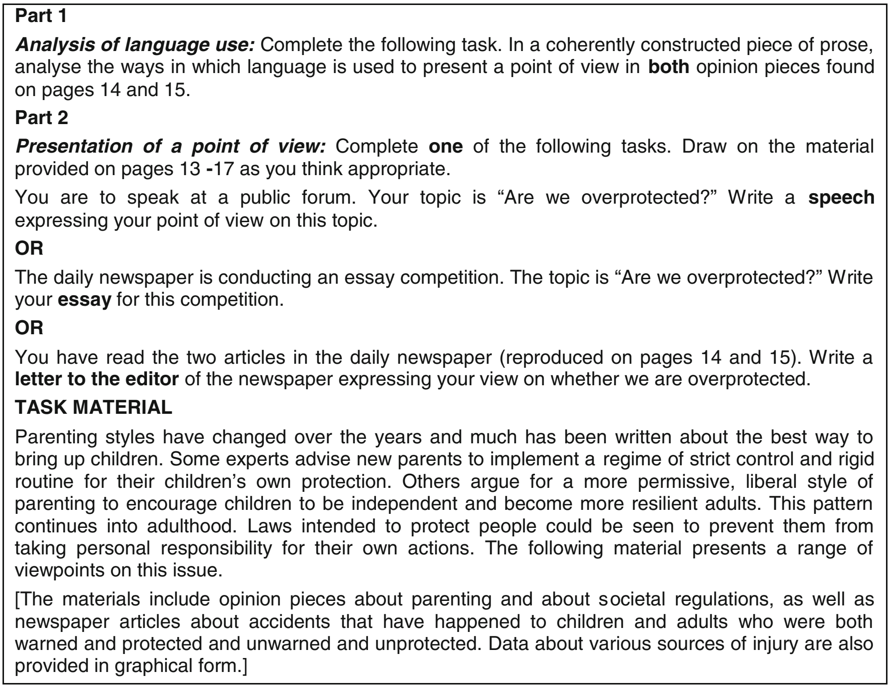
Victoria. In Victoria, as in many other Australian states, a mixed system of centralized and decentralized assessment combines these kinds of school-based assessment practices with a set of state exams guided by the Victoria Essential Learning Standards (VELS). Considerable attention is given to teachers' abilities to assess the VELS. The standards define what students should know and be able to do at each level so that units of work based on activities described in the learning focus statements are assessable against the expected standards. An emphasis on real-world tasks supports transfer in learning. Assessment maps are provided within each domain to assist teachers in assessing all the standards. These are a collection of student work samples for each domain, each of which is annotated to describe attributes of the student's work and its relationship with specific elements of the standards, as well as progression points illustrating development within each level. Teachers are advised that:

> Assessment of student achievement against the standards requires a mix of summative assessment to determine what the student has achieved and formative assessment to inform the next stage of learning. This should be based on authentic assessment in which students are asked to perform real-world tasks demonstrating the application of essential knowledge and skill. Assessment must **also** evaluate knowledge, skills and behaviours in an integrated way, rather than treating each and every standard as discrete. This not only ensures a more efficient approach to student assessment that avoids unnecessary duplication of assessment tasks and subsequent reports, but also more clearly reflects how students actually learn and develops deep understanding in learners which can be transferred to new and different contexts (VCAA 2009).

At the secondary level, the *Victorian Certificate of Education (VCE)* provides guide pathways to further study at university, Technical and Further Education (TAFE) and to the world of work. Some students undertake a school-based apprenticeship or traineeship within the VCE. The Victoria Curriculum and Assessment Authority establishes courses in a wide range of studies, develops the external examinations, and ensures the quality of the school-assessed component of the VCE.

VCAA conceptualizes assessment as "of," "for," and "as" learning. Teachers are involved in developing assessments, along with university faculty in the subject area, and all prior year assessments are public, in an attempt to make the standards and means of measuring them as transparent as possible. Before the external examinations are given to students, teachers and academics sit and take the exams themselves, as though they were students. The external subject-specific examinations, given in grades 11 and 12, include about 25% machine-scored items; the remaining items are open-ended and are scored by the classroom teacher. The exams may include written, oral, and performance elements. Language examinations, for example, include on-demand oral tests, and arts examinations include required performance components, such as dance and musical performances.

The VCE exams often push toward applications of knowledge in problem-solving contexts requiring evaluation and innovative thinking. For example, the Design and Technology exam poses several design challenges to which students have to respond along many dimensions—with respect to materials, engineering features, safety, reliability, and aesthetic considerations—while resolving design dilemmas and justifying their decisions.



Part 1

Analysis of language use: Complete the following task. In a coherently constructed piece of prose, analyse the ways in which language is used to present a point of view in **both** opinion pieces found on pages 14 and 15.

Part 2

Presentation of a point of view: Complete **one** of the following tasks. Draw on the material provided on pages 13 **-**17 as you think appropriate.

You are to speak at a public forum. Your topic is “Are we overprotected?” Write a **speech** expressing your point of view on this topic.

OR

The daily newspaper is conducting an essay competition. The topic is “Are we overprotected?” Write your **essay** for this competition.

OR

You have read the two articles in the daily newspaper (reproduced on pages 14 and 15). Write a **letter to the editor** of the newspaper expressing your view on whether we are overprotected.

TASK MATERIAL

Parenting styles have changed over the years and much has been written about the best way to bring up children. Some experts advise new parents to implement a regime of strict control and rigid routine for their children’s own protection. Others argue for a more permissive, liberal style of parenting to encourage children to be independent and become more resilient adults. This pattern continues into adulthood. Laws intended to protect people could be seen to prevent them from taking personal responsibility for their own actions. The following material presents a range of viewpoints on this issue.

[The materials include opinion pieces about parenting and about societal regulations, as well as newspaper articles about accidents that have happened to children and adults who were both warned and protected and unwarned and unprotected. Data about various sources of injury are also provided in graphical form.]

Fig. 6.7 High school english examination question, Victoria, Australia

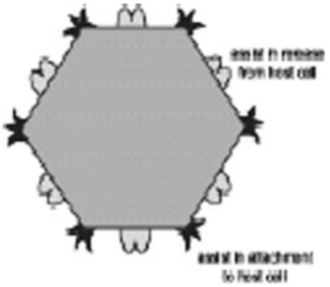
In the on-demand portion of the English exam, which is comprised of several essays that test aspects of analysis and communication skills, students must analyze aspects of literature they have read, respond to critical interpretations of texts with their own analyses and ideas, and develop and explain their thinking about a topic after reading several source materials that provide differing kinds of information and points of view. In one such task, students are asked to analyze whether parents and government laws seek to “overprotect” citizens from potential harm (see Fig. 6.7).

In addition to the on-demand tests, at least 50% of the total examination score is comprised of classroom-based tasks that are given throughout the school year. Teachers design these required assignments and assessments—lab experiments and investigations on central topics as well as research papers and presentations—in response to syllabus expectations. These required classroom tasks ensure that students are getting the kinds of learning opportunities which prepare them for the assessments they will later take, that they are getting the feedback that they need to improve, and that they will be prepared to succeed, not only on these very challenging tests but also at college and in life, where they will have to apply knowledge in these ways.

When scientists design drugs against infectious agents, the term "designer drug" is often used.

Explain what is meant by this term.

Scientists aim to develop a drug against a particular virus that infects humans. The virus has a protein coat and different parts of the coat play different roles in the infective cycle. Some sites assist in the attachment of the virus to a host cell; others are important in the release from a host cell. The structure is represented in the following diagram:


assist in release from host cell
assist in attachment to host cell

The virus reproduces by attaching itself to the surface of a host cell and injecting its DNA into the host cell. The viral DNA then uses the components of host cell to reproduce its parts and hundreds of new viruses bud off from the host cell. Ultimately the host cell dies.

Design a drug that will be effective against this virus. In your answer outline the important aspects you would need to consider. Outline how your drug would prevent continuation of the cycle of reproduction of the virus particle. Use diagrams in your answer. Space for diagrams is provided on the next page.

Before a drug is used on humans, it is usually tested on animals. In this case, the virus under investigation also infects mice. Design an experiment, using mice, to test the effectiveness of the drug you have designed.

Fig. 6.8 High school biology examination question, Victoria, Australia

An example from the Victoria biology test, shown in Fig. 6.8, describes a particular virus to students, asks them to design a drug to kill the virus and, in several pages, to explain how the drug operates, and then to design an experiment to test it.

In preparation for this on-demand test, students taking Biology will have been assessed on six pieces of work during the school year covering specific outcomes in the syllabus. For example, they will have conducted "practical tasks" such as using a microscope to study plant and animal cells by preparing slides of cells, staining them, and comparing them in a variety of ways, resulting in a written product with visual elements. They also will have conducted practical tasks on enzymes and membranes, and on the maintenance of stable internal environments for animals and plants. Finally, they will have completed and presented a research report on characteristics of pathogenic organisms and mechanisms by which organisms can defend against disease. These tasks, evaluated as part of the final examination score, link directly to the expectations that students will encounter on the external examination, but go well beyond what that examination can measure in terms of how students can apply their knowledge.

The tasks are graded according to the criteria set out in the syllabus. The quality of the tasks assigned by teachers, the work done by students, and the appropriateness of the grades and feedback given to students are audited through an inspection system, and schools are given feedback on all of these elements. In addition, the VCAA uses statistical moderation to ensure that the same assessment standards are applied to students across schools. The external exams are used as the basis for this moderation, which adjusts the level and spread of each school's assessments of its students to match the level and spread of the same students' collective scores on the common external test score. The system supports a rich curriculum and ambitious assessments for students with a comparable means for examining student learning outcomes.

Finland

Finland has been much studied since it climbed rapidly, over a decade and a half, to the top of the international rankings for both economic competitiveness and educational outcomes. In 2006, it ranked first among the OECD nations on the PISA assessments in mathematics, science, and reading. Leaders in Finland attribute these gains to their intensive investments in teacher education and major overhaul of the curriculum and assessment system (Laukkanen 2008; Buchberger and Buchberger 2004). Prospective teachers are competitively selected from the pool of college graduates and receive a 3-year graduate-level teacher preparation program, entirely free of charge and with a living stipend. Their master's degree program offers a dual focus on inquiry-oriented teaching and teaching that meets the needs of diverse learners—and includes at least a full year of clinical experience in a model school associated with the university. Preparation includes a strong focus on how to use formative performance assessments in the service of student learning.

Policy makers decided that if they invested in very skillful teachers, they could allow local schools more autonomy to decide what and how to teach—a reaction against the highly centralized system they sought to overhaul. Finland's national core curriculum is a much leaner document, reduced from hundreds of pages of highly specific prescriptions to descriptions of a small number of skills and core concepts each year (e.g., the full set of math standards for all grades are described in about ten pages). This guides teachers in collectively developing local curricula and assessments that encourage students to be active learners who can find, analyze, and use information to solve problems in novel situations.

There are no external standardized tests used to rank students or schools. Finland's leaders point to the use of school-based, student-centered, open-ended tasks embedded in the curriculum as an important reason for the nation's extraordinary success on international examinations (Lavonen 2008; Finnish National Board of Education 2007). Finnish education authorities periodically evaluate school-level samples of student performance, generally at the end of the 2nd and 9th grades, to inform curriculum and school investments. All other assessments are designed and managed locally. The national core curriculum provides teachers with recommended assessment criteria for specific grades in each subject and in the overall final assessment of student progress each year (Finnish National Board of Education June 2008).

Local schools and teachers then use those guidelines to craft a more detailed curriculum and set of learning outcomes at each school as well as approaches to assessing benchmarks in the curriculum (Finnish National Board of Education June 2008). Teachers are treated as "pedagogical experts" who have extensive decision-making authority in the areas of curriculum and assessment as in other areas of school policy and management (Finnish National Board of Education April 2008).

According to the Finnish National Board of Education (June 2008), the main purpose of assessing students is to guide and encourage students' own reflection and self-assessment. Consequently, on-going feedback from the teacher is very important. Teachers give students formative and summative reports both through verbal feedback and on a numerical scale based on students' level of performance in relation to the objectives of the curriculum. All Finnish schools use a grading scale of 4–10, where 5 is "adequate" and 10 is "excellent." The recommended assessment criteria are shaped around the grade of 8 or "good." Teachers' reports must be based on multiple forms of assessment, not just exams. Schools are responsible for giving basic education certificates for completing the different milestones of comprehensive school up to 9th grade and additional classes prior to university (European Commission 2007/2008).

Most Finnish students take a set of voluntary matriculation examinations that provide information for university admissions based on students' abilities to apply problem-solving, analytic, and writing skills. University and high school faculty members construct the examinations—which are composed of open-ended essays and problem solutions—under the guidance of the Matriculation Exam Board, which is appointed by the Finnish Ministry of Education to organize, manage, and administer the exam (The Finnish Matriculation Examination 2008). The board members (about 40 in number) are faculty and curriculum experts in the subject areas tested, nominated by universities and the National Board of Education. More than 300 associate members—also typically high school and college faculty—help develop and review the tests. High school teachers grade the matriculation exams locally using official guidelines, and samples of the grades are reexamined by professional raters hired by the board (Kaftandjieva and Takala 2002).

Students take at least four exams, with a test in the students' mother tongue (Finnish, Swedish, or Saami) being compulsory. These tests have a textual skills section that evaluates students' analytic skills and linguistic expression, and an essay that focuses on the development of thinking, linguistic expression, and coherence. They then choose three other tests from among the following: the test in the second national language, a foreign language test, the mathematics test, and one or more tests from the general battery of tests in the sciences and humanities (e.g., religion, ethics, philosophy, psychology, history, social studies, physics, chemistry, biology, geography, and health education). The tests also incorporate questions which cross disciplinary boundaries.

The Finnish system assumes that all students aiming for college (who comprise a majority) will be at least bilingual and that many will be trilingual. The language tests evaluate listening and reading comprehension as well as writing in the language in question.

In addition to choosing which tests to take, students make choices of which items to answer within the exams. In the general battery, they are generally given a set of questions or prompts from which they must respond to six or eight of their choice. On the mathematics test, there are 15 or so problems from which they must choose 10 to answer. Problems require critical thinking and modeling, as well as straightforward problem-solving.

For example, the Basic Mathematics exam poses this problem:

> A solution of salt and water contains 25 per cent salt. Diluted solutions are obtained by adding water. How much water must be added to one kilogram of the original solution in order to obtain a 10 per cent solution? Work out a graphic representation which gives the amount of water to be added in order to get a solution with 2–25% of salt. The amount of water (in kilograms) to be added to one kilogram of the original solution must be on the horizontal axis; the salt content of the new solution as a percentage must be on the vertical axis.

And the Advanced Mathematics exam poses this one:

> In a society the growth of the standard of living is inversely proportional to the standard of living already gained, i.e. the higher the standard of living is, the less willingness there is to raise it further. Form a differential-equation-based model describing the standard of living and solve it. Does the standard of living rise forever? Is the rate of change increasing or decreasing? Does the standard of living approach some constant level?

Assessment is used in Finland to cultivate students' active learning skills by posing complex problems and helping students address these problems. For example, in a Finnish classroom, it is rare to see a teacher standing at the front of a classroom lecturing students for 50 minutes. Instead, teachers are likely to be coaching students who are working on hands-on tasks that are often self-managed. A description of a Finnish school (Korpela 2004) illustrates how students may be engaged in active, self-directed learning, rotating through workshops or gathering information, asking questions of their teacher, and working with other students in small groups. They may be focusing on completing independent or group projects or writing articles for their own magazine. The cultivation of independence and active learning allows students to focus on broad knowledge with emphasis on skills like analytical thinking, problem-solving, and metacognitive skills that develop students' thinking (Lavonen 2008).

Although not part of the mandatory national assessment system, one assessment project of some potential interest to ATC21S is the "Learning to Learn" project launched in the mid-1990s as a partnership between the Finnish National Board of Education, the Centre for Educational Assessment at the University of Helsinki, and the City of Helsinki Education Department. Reports through 2002 describe the results of several studies of 6th grade, 9th grade, and upper secondary school students using cognitive and affective measures administered as paper-and-pencil test items and attitudinal surveys (Hautamaki et al. 2002; Hautamaki and Kupiainen 2002). The project developed an elaborated framework for conceptualizing "learning to learn," defining it in the summary report as:

> … the adaptive and voluntary mastery of learning action. After initial task acceptance, learning action is seen to be maintained through affective and cognitive self-regulation.

> Learning-to-learn can then be defined as the readiness and willingness to adapt to a novel task. It consists of a complex system of cognitive competencies and self- and context-related beliefs. Readiness, or cognitive competence, refers both to the knowledge of relevant facts and to the use of thinking and reasoning; i.e., to the retrieval of the already learnt and to the application of general procedures to adapt to new situations. The cognitive component of learning-to-learn is also referred to as *mastery of reasoning*. It is related to Piaget's reflective abstraction, and the scaling of the indicator is criterion-referenced in relation to the mastery of formal operational schemata. This distinguishes it from classical measures of intelligence, as concrete and formal operations can be shown to be malleable and thus teachable. The affective component of learning-to-learn is seen to consist of several relatively independent subsystems, comprising both self- and context-related beliefs. Among these, learning motivation, action-control beliefs, school-subject-related beliefs, task acceptance, socio-moral commitment, self-evaluation, and the experienced support of significant others are seen to be central when learning-to-learn is assessed at school level (Hautamäki and Kupiainen 2002, pp. 3–4).

That report noted both the interest generated by this conceptual framework (for a full discussion, see Hautamaki et al. 2002), along with some concerns about the assessment formats, in particular the use of paper-and-pencil, multiple-choice items in collecting data. The researcher observed that "The 'real' learning situations in later life are not in a ready paper-and-pencil form" (Hautamaki and Kupiainen 2002, p. 22) and suggested that further work on open-ended prompts and real-life tasks (coming nearer to a work-sample approach) would be closer to ideal if cost considerations could be overcome.

Singapore

In Singapore more recently, greater emphasis has been placed on school-based assessment integrated into large-scale testing systems. Singapore's education system has been a source of intense interest for policy analysts since its students took first place in the TIMSS (Trends in International Mathematics and Science Study) assessments in mathematics and science in 1995, 1999, and 2003. These rankings are based on strong achievement for all of the country's students, including the Malay and Tamil minorities, who have been rapidly closing what was once a yawning achievement gap (Dixon 2005). About 90% of Singapore's students scored above the international median on the TIMSS tests. This accomplishment is even more remarkable, given that fewer than half of Singapore's students routinely speak English, the language of the test, at home. Most speak one of the other official national languages of the country—Mandarin, Malay, or Tamil—and some speak one of several dozen other languages or dialects.

Intensive investment and reform over 30 years have transformed the Singaporean education system, broadening access and increasing equality, while orchestrating a system that includes a complex system of private, "autonomous," and public schools, some of them inherited from the colonial era, all of which receive government subsidies. These schools are intentionally diverse in many ways, as local schools are urged to innovate, but purposely have common instructional expectations and supports, with a common national curriculum for core subjects.

Since the prime minister introduced the “thinking schools, learning nation” initiative in 1997, Singapore’s explicit focus within its reforms of curriculum, assessment, and teaching has been to develop a creative and critical thinking culture within schools by explicitly teaching and assessing these skills for students—and by creating an inquiry culture among teachers as well, who are given support to conduct action research on their teaching and continually to revise their teaching strategies in response to what they learn. This initiative has been married to a commitment to integrate technology into all aspects of education—a mission nearly fully accomplished a decade later—and to open up college and university admissions dramatically.

Higher education is now available to virtually every Singaporean. Based on their interests, labor force needs, and the results of their grades, O-level exams, and other accomplishments, students pursue one of three pathways after 10th grade, when secondary school ends: about 25% attend Junior College for 2 years, followed by university, which leads to professional paths such as teaching, science, engineering, medicine, law, and the civil service; about 60% attend a polytechnic college for 3 years, after which about half go on to the university while the others go into jobs in technical and engineering fields; and the remainder—about 15%—attend an Institute of Technical Education for 2 years, and, even then, some continue onto college or university. Virtually everyone finishes one of these pathways.

Historically, the schools have operated a modified British-style system. Students sit for national exams administered by the Singapore Examinations and Assessment Board (SEAB). At the end of year 6 (age 12), students take the Primary School Leaving Examinations (PSLE), which are open-ended written and oral examinations in four core subject areas: mathematics, science, English, and a “mother tongue” language, administered and scored by teachers in moderated scoring sessions. The exams in English and native languages include four components—two written essays of at least 150 words, listening comprehension, language comprehension, and an oral exam that requires students to engage in a conversation on a set topic for 15 min. Two examiners observe the candidates and grade the oral proficiency of the student. In mathematics, students have to demonstrate the steps in solving a problem.

Students then take the General Certificate of Examinations Normal or Ordinary Level (GCE N/O-Level) at the end of year 10 (age 16). The GCE N- and O-level examinations are based on common course syllabi that outline what is to be taught; they require short and long open-ended responses and essays across a wide range of content areas from which students choose the ones in which they want to be examined. Although the results are used to guide postsecondary admissions, not to determine graduation from high school, they exert substantial influence on the high school curriculum. Recent reforms are changing the curriculum and assessment system to make it more explicitly focused on creativity and independent problem-solving. Many courses include applied examination elements that allow students to demonstrate how they can solve problems in performance tasks.

For example, the examination score for the Computer Applications course at N-level includes a paper and pencil component (30%), a practical component (35%), and a specific set of course-embedded tasks (35%) to be scored by teachers using common criteria. The practical examination tests students’ ability to use both word processing and spreadsheet software for a series of tasks. The course-embedded project requires students to design a database, website, or product using technology. At O-level, the Computer Applications exam requires a school-based project (25%) that runs over a 14-week period. Students must identify a problem they want to tackle, design a technology-based solution, implement the solution, design and implement a testing strategy to evaluate it, document their strategy and the results of their testing, and evaluate the success and limitations of the overall solution strategy. These examination elements are scored by teachers using common criteria with internal and external moderation of scores for comparability.

Students attending Junior College (grades 11 and 12) en route to university take the GCE Advanced Level (A-Level) exams at the end of year 12 (age 18). A new "A"-level curriculum and examination system was introduced in 2002. The new exams are meant to encourage multidisciplinary learning by requiring that students "select and draw together knowledge and skills they have learned from across different subject areas, and apply them to tackle new and unfamiliar areas or problems" (Singapore Examinations and Assessment Board 2006, p. 2).

The A-level curricular framework includes core content areas in which students take courses and associated exams: humanities, mathematics, sciences, and languages. It also includes Life Skills—emphasizing leadership, enrichment, and service to others—and Knowledge Skills, evaluated through a general paper, project work, and a course in knowledge and inquiry. A typical A-level student is evaluated in three compulsory subjects—a general paper, project work, and a native language assessment—along with four content subjects.

The newer areas of Life Skills and Knowledge Skills are intended to develop the more advanced thinking skills thought to be underrepresented in the traditional content-based curriculum and examination system. They represent the goals of reforms launched in 1997 as part of the "thinking schools, learning nation" initiative, which created a number of changes:

Syllabi, examinations and university admission criteria were changed to encourage thinking out of the box and risk-taking. Students are now more engaged in project work and higher order thinking questions to encourage creativity, independent, and inter-dependent learning (Ng 2008, p. 6).

The content courses are also evolving to include more critical thinking, inquiry, and investigation, along with mastery of content. A number of the high school content tests are accompanied by school-based tasks, such as research projects and experiments designed and conducted by students. Each of the science courses now includes a component called the "School-based Science Practical Assessment" (SPA). These school-based components, managed and scored by teachers according to specifications provided by the Examinations Board, count for up to 20% of the examination grade. Scoring is both internally and externally moderated. The goal is for students to be able to:

1. Follow a detailed set or sequence of instructions and use techniques, apparatus, and materials safely and effectively
2. Make and record observations, measurements, methods, and techniques with precision and accuracy

3. Interpret and evaluate observations and experimental data
4. Identify a problem, design and plan investigations, evaluate methods and techniques, and suggest possible improvements in the design

The projects can be submitted to the university as part of the application, and universities are encouraged to examine evidence about student accomplishments beyond examination scores. Below we describe some of these innovations in the examination system.

Innovative Features of the Examination System

Project Work

Project work (PW) is *an interdisciplinary subject* that is compulsory for all pre-university students. There is dedicated curriculum time for students to carry out their project tasks over an extended period. As an interdisciplinary subject, it breaks away from the compartmentalization of knowledge and skills to focus on interdisciplinary outcomes by requiring students to draw knowledge and apply skills from across different subject domains. The goals for this experience are embedded in the requirements for the task and its assessment, which are centrally set by the Singapore Examinations and Assessment Board. The tasks are designed to be sufficiently broad to allow students to carry out a project that they are interested in while meeting the task requirements:

- *It must foster collaborative learning through group work*: Together, as a group randomly formed by the teacher, students brainstorm and evaluate each others' ideas, agree on the project that the group will undertake, and decide on how the work should be allocated among themselves.
- *Every student must make an oral presentation*: Individually and together as a group, each student makes an oral presentation of his/her group project in front of an audience.
- Both product and process are assessed: There are 3 components for assessment:
  - The *Written Report* which shows evidence of the group's ability to generate, analyze, and evaluate ideas for the project.
  - The *Oral Presentation* in which each individual group member is assessed on his/her fluency and clarity of speech, awareness of audience as well as response to questions. The group as a whole is also assessed in terms of the effectiveness of the overall presentation.
  - The *Group Project File* in which each individual group member submits three documents related to "snapshots" of the processes involved in carrying out the project. These documents show the individual student's ability to generate, analyze, and evaluate (1) preliminary ideas for a project, (2) a piece of research material gathered for the chosen project, and (3) insights and reflections on the project.

In carrying out the PW assessment task, students are intended to acquire self-directed inquiry skills as they propose their own topic, plan their timelines, allocate individual areas of work, interact with teammates of different abilities and personalities, and gather and evaluate primary and secondary research material. These PW processes reflect life skills and competencies, such as knowledge application, collaboration, communication, and independent learning, which prepare students for the future workplace.

About 12,000 students complete this task annually. Assessment is school-based and criterion-referenced. While task setting, conditions, assessment criteria, achievement standards, and marking processes are externally specified by SEAB, the assessment of all three components of PW is carried out by classroom teachers, using a set of assessment criteria provided by the board. All schools are given exemplar material that illustrates the expected marking standards. The board provides training for assessors and internal moderators. Like all other assessments, the grading is both internally and externally moderated.

Knowledge and Inquiry

Knowledge and inquiry is a Humanities subject that seeks to develop in students:

- *An understanding of the nature and construction of knowledge*: Students are expected to show that they have read widely and have understood and can apply the concepts involved. They are expected to demonstrate skill in selecting relevant material with which to tackle the assessment tasks.
- *Critical thinking*: Students are expected to demonstrate skills of critical thinking. They are expected to analyze different kinds of arguments and information, identify and evaluate assumptions and points of view, verify claims, and provide reasoned and supported arguments of their own.
- *Communication skills*: Students are expected to communicate their ideas and arguments clearly and coherently in good English. They are expected to structure their arguments and to select an appropriate style of presentation, to communicate responses which are fully relevant to the questions asked and to demonstrate a clear ability to engage with different aspects of these questions.

There are three assessment components:

- *Essay*: This paper gives candidates the opportunity to demonstrate their ability to apply the concepts they have learned in their study of the nature and construction of knowledge. It covers the theoretical aspects of areas of exploration identified in the syllabus, and the questions set will require candidates to draw on knowledge they have gained during their study of the following key questions:

  - Why ask questions?
  - What is knowledge?
  - How is knowledge constructed?
  - What makes knowledge valid?

  - How is knowledge affected by society?
  - How should knowledge be used?

- *Critical thinking*: This paper requires students to critically analyze different kinds of arguments and information presented in the material, identify and evaluate assumptions and points of view, and verify claims, and to provide reasoned and supported arguments. Students must use language appropriately and effectively to communicate a clear and well-structured argument.
- *Independent study*: The independent study component allows students to demonstrate their understanding of the nature and construction of knowledge as it relates to their chosen area of study, apply this understanding in addressing the specific context, select appropriate material, and show that they have engaged in relevant reading during the course of their research by presenting a literature review and applying what they have read to support the arguments they present. Students must use language appropriately and effectively to communicate a clear and well-structured argument. At the end of the 6 months of independent research study, they submit an extended essay of 2,500–3,000 words.

The kinds of more intellectually challenging school-based assessment in the high school examinations are also encouraged in the earlier grades as well. In the curriculum and assessment guidelines that accompany the national standards, teachers are encouraged to engage in continual assessment in the classroom, using a variety of assessment modes, such as classroom observations, oral communication, written assignments and tests, and practical and investigative tasks. The Ministry has developed a number of curriculum and assessment supports for teachers. For example, SAIL (Strategies for Active and Independent Learning) aims to support more learner-centered project work in classrooms and provides assessment rubrics to clarify learning expectations. All schools have received training in using these tools.

The Ministry's 2004 Assessment Guides for both primary and lower secondary mathematics contain resources, tools, and ideas to help teachers incorporate strategies such as mathematical investigations, journal writing, classroom observation, self-assessment, and portfolio assessment into the classroom. Emphasis is placed on the assessment of problem-solving and on metacognition, the self-regulation of learning that will enable students to internalize standards and become independent learners (Kaur 2005). The Institute of Education has held a variety of workshops to support learning about the new assessments and integrated the new strategies into teacher development programs.

United Kingdom

The move toward more school-based assessment has also occurred in various ways in the UK, which, for more than a century, has had some influence on examination systems in English-speaking countries around the world. Assessments have typically been open-ended essay and constructed-response examinations, but the nature of the tasks and the form of administration have been changing over the last two decades to include more school-based tasks and projects.

England

England's assessment system is managed at the national level by an organization called the Qualifications and Curriculum Authority (QCA). Schools teach and assess students using a national curriculum, which includes syllabi for specific courses. Teachers assess pupils' progress continuously and assemble evidence for external reporting in the national data system at ages 7, 11, and 14 (key stages 1, 2, and 3). This evidence is based on classroom-based assignments, observations, and tasks, the results of which are evaluated in terms of indicators of performance outlined in learning progressions for each of several dimensions of learning within each subject area.

At key stage 1, ages six to seven, student progress is evaluated on the basis of classroom evidence and results from centrally developed, open-ended tests and tasks in English and mathematics. These tests and tasks are marked by teachers and moderated within the school and by external moderators. At key stage 2, ages 8 through 11, student progress is evaluated on the basis of teachers' summary judgments and results from open-ended tests in English, mathematics, and science. These tests are externally marked and the results reported on a national level. At key stage 3, England has recently abolished external tests and now relies on teacher assessments to report achievement levels in all subjects. Teacher judgments are moderated, and results are reported on a national level.

The Assessing Pupils' Progress program that guides this work is described by the QCA in this way:

APP is the new structured approach to teacher assessment, developed by QCA in partnership with the National Strategies, which equips teachers to make judgments on pupils' progress. It helps teachers to fine-tune their understanding of learners' needs and to tailor their planning and teaching accordingly, by enabling them to: use diagnostic information about pupils' strengths and weaknesses to improve teaching, learning and pupils' progress; make reliable judgments related to national standards drawing on a wide range of evidence; and track pupils' progress.

The APP subject materials for teachers include assessment guidelines for assessing pupils' work in relation to national curriculum levels. These provide a simple recording format providing assessment criteria for each of the assessment focuses in the subject, and standards files, which are annotated collections of pupils' day-to-day work that exemplify national standards at different levels. These help teachers reach consistent and reliable judgments about national curriculum levels (Qualifications and Curriculum Authority 2009, p. 1.)

Some nationally developed tasks are designed and distributed to schools to support teacher assessment. At key stage 2 (age 11), a set of these tasks and tests must be used to evaluate students, in combination with the other evidence teachers assemble from the classroom. In other years, the use of the tasks is optional. As described by the QCA: “The tasks are designed to support teacher assessment. They can be used to indicate what pupils are able to do and inform future learning and teaching strategies. Individual tasks can be used to provide a basis for discussion by teachers and pupils on what has been achieved and to identify the next steps. They can support day-to-day assessment and generate outcomes which can contribute to the breadth of evidence which is used as the basis for periodic and transitional assessment.”

At key stage 4, ages 15–16, the national qualification framework includes multiple pathways for students and consequently multiple measures of student achievement. There are four pathways based on students’ aspirations after graduation: apprenticeship, diploma, the General Certificate of Secondary Education (GCSE), and the A-Level examinations. Some students go on to a Further Education college to take vocationally related courses. They usually take the National Vocational Qualification using the apprenticeship model.

Most students take the GCSE, a 2-year course of study evaluated by assessments both within and at the end of courses or unit. Students may take as many single-subject or combined-subject assessments as they like, and they choose which ones to take on the basis of their interests and areas of expertise. The exams involve constructed response items and structured, extended classroom-based tasks, which comprise from 25% to 60% of the final examination score. England is currently piloting new tasks for the GCSE with an increased emphasis on functional skills like problem-solving, team building, and communication as well as personal learning and thinking skills across subjects. These new tasks, called “controlled assessments” are either designed by the awarding body and marked by teachers or designed by teachers and marked by the awarding body. Either way teachers determine the timing of controlled assessments.

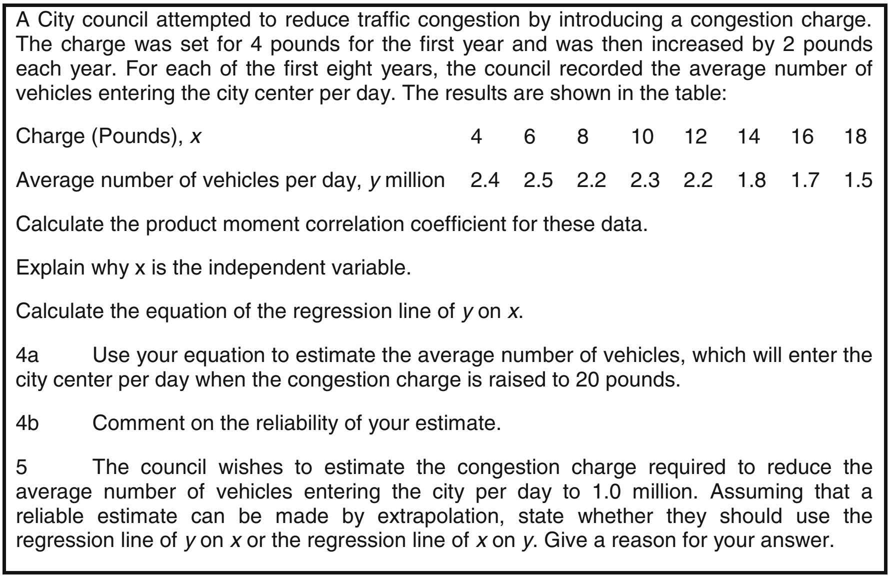
These classroom-based assessments comprise 25% of the total examination score in subjects like business studies, classical civilization, English literature, geography, history, humanities, or statistics, and 60% of the total examination score in subject areas such as applied business, music and dance, design and technology, drama, engineering, English, English Language, expressive arts, health and social care, home economics, ICT, manufacturing, media studies, and modern foreign languages. Examples of classroom-based tasks in English are given in Table 6.2 and in Interactive Computer Technology (ICT) in Fig. 6.9.

During key stage 4, most students take five or more GCSE exams. Their performance determines the level of the diploma they receive, and whether they will go on to Advanced Studies that are later evaluated by A-level exams that qualify students for university admissions. England has 45 areas for A-level exams. The exam questions require extended answers aimed at assessing deeper levels of understanding and applications of knowledge to real-world problems, as illustrated in the example in Fig. 6.10.

Most of the exams take the form of essay questions. The mathematics exams include questions that ask students to show their reasoning behind their answers.

Table 6.2 Classroom-based assessment tasks, english GCSE

Unit and assessment	Tasks
Reading literacy texts controlled assessment (coursework) 40 marks	Responses to three texts from choice of tasks and texts. Candidates must show an understanding of texts in their social, cultural, and historical context
Imaginative writing controlled assessment (coursework)40 marks	Two linked continuous writing responses from a choice of Text Development or Media
Speaking and listening controlled assessment (coursework) 40 marks	Three activities: a drama-focused activity, a group activity, an individual extended contribution. One activity must be a real-life context in and beyond the classroom
Information and ideas written examination 80 marks (40 per section)	Nonfiction and media: Responses to unseen authentic passages Writing information and ideas: One continuous writing response—choice from 2 options



A City council attempted to reduce traffic congestion by introducing a congestion charge. The charge was set for 4 pounds for the first year and was then increased by 2 pounds each year. For each of the first eight years, the council recorded the average number of vehicles entering the city center per day. The results are shown in the table:

Charge (Pounds), *x*	4	6	8	10	12	14	16	18
Average number of vehicles per day, *y* million	2.4	2.5	2.2	2.3	2.2	1.8	1.7	1.5

Calculate the product moment correlation coefficient for these data.

Explain why x is the independent variable.

Calculate the equation of the regression line of *y* on *x*.

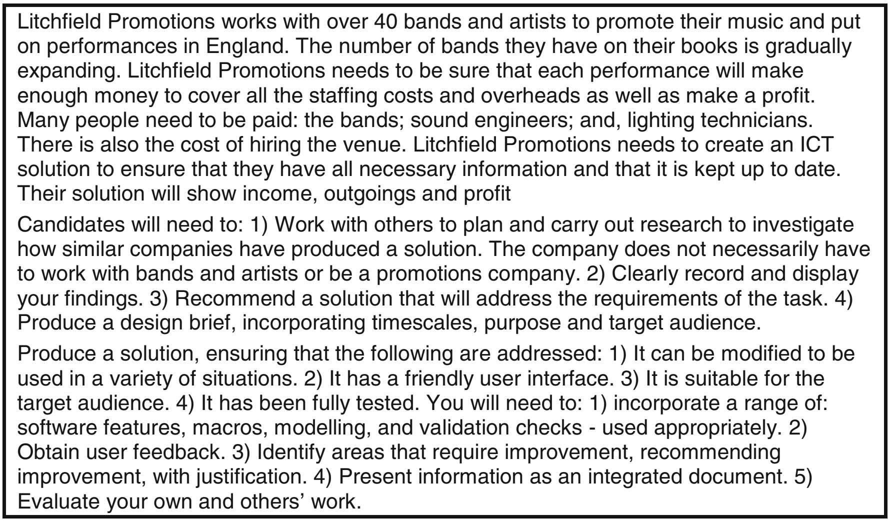
4a Use your equation to estimate the average number of vehicles, which will enter the city center per day when the congestion charge is raised to 20 pounds.

4b Comment on the reliability of your estimate.

5 The council wishes to estimate the congestion charge required to reduce the average number of vehicles entering the city per day to 1.0 million. Assuming that a reliable estimate can be made by extrapolation, state whether they should use the regression line of *y* on *x* or the regression line of *x* on *y*. Give a reason for your answer.

Fig. 6.9 English A-level question from a probability and statistics examination

Foreign language exams require oral presentations. The "A"-level exam in English literature asks students to show their skills and knowledge in four sections: poetry, drama, prose, and general, analyzing works of literature they have read as part of their curriculum in terms of their meaning and interpretation as well as literary devices and writing strategies. Coursework accounts for 25–30% of the "A"-level score, depending on the course. Students must now also complete an independently



Litchfield Promotions works with over 40 bands and artists to promote their music and put on performances in England. The number of bands they have on their books is gradually expanding. Litchfield Promotions needs to be sure that each performance will make enough money to cover all the staffing costs and overheads as well as make a profit. Many people need to be paid: the bands; sound engineers; and, lighting technicians. There is also the cost of hiring the venue. Litchfield Promotions needs to create an ICT solution to ensure that they have all necessary information and that it is kept up to date. Their solution will show income, outgoings and profit

Candidates will need to: 1) Work with others to plan and carry out research to investigate how similar companies have produced a solution. The company does not necessarily have to work with bands and artists or be a promotions company. 2) Clearly record and display your findings. 3) Recommend a solution that will address the requirements of the task. 4) Produce a design brief, incorporating timescales, purpose and target audience.

Produce a solution, ensuring that the following are addressed: 1) It can be modified to be used in a variety of situations. 2) It has a friendly user interface. 3) It is suitable for the target audience. 4) It has been fully tested. You will need to: 1) incorporate a range of: software features, macros, modelling, and validation checks - used appropriately. 2) Obtain user feedback. 3) Identify areas that require improvement, recommending improvement, with justification. 4) Present information as an integrated document. 5) Evaluate your own and others' work.

Fig. 6.10 Controlled assessment tasks, interactive computer technology GCSE

designed extended research project as part of the A-level assessments. Teachers mark assessments in a moderated process managed by the five examination agencies that organize sets of examinations.

While England has moved to include some school-based assessments in its increasingly performance-oriented assessment system, Scotland, Wales, and Northern Ireland have gone even further in revising their approaches to assessment.

Scotland

Scotland has a governing body for its educational system that is separate from that of the UK and uses a set of assessments called the Scottish Survey of Achievement, administered in the third, fifth, and seventh years of primary school as well as standardized courses and benchmark exams in secondary school. The assessment tasks for the primary courses and general secondary courses are designed and marked by teachers and lecturers. Schools use external assessments for the intermediate and advanced secondary courses. The Scottish Qualifications Authority designs and scores those assessments which may take the form of examinations, project work, or portfolios (Scottish Qualifications Authority March 2004; The Scottish Government 2008).

Wales

Wales recently separated from the system used in England and now has its own governing body for its educational system (Archer 2006). Wales abolished national exams for children through age 14. Much like Finland, during the primary years, Welsh schools have a national school curriculum supported by teacher-created, administered, and scored assessments. During the secondary years, teachers create and manage all assessment of 14-year-old students, while students 16 years and older are encouraged to participate in the relevant GCSE exams and A-level courses and exams administered by the U.K.'s Qualifications and Curriculum Authority (Welsh Assembly Government 2008a, b). With these changes to its assessment system, Wales hopes to increase student engagement, engage students in more creative tasks, and reduce teaching to the test (Archer 2006).

Northern Ireland

Northern Ireland is in the process of implementing an approach at all levels called "Assessment for Learning." This approach emphasizes locally developed, administered, and scored assessments and focuses on five key actions:

1. *Sharing learning intentions* where students and teacher agree upon learning intentions to give them ownership over their learning.
2. *Sharing and negotiating success criteria* where students and teacher create the criteria for successful completion of a task together to help with self-assessment.
3. *Feedback* where teachers provide on-going feedback during formative assessment sessions.
4. *Effective questioning* where teachers introduce strategies like using open-ended questions and giving more thinking time so students will feel more confident thinking aloud and explaining their reasoning.
5. *How pupils reflect on their learning* where teachers provide students with strategies to think about what they have learned.

Northern Ireland does not require schools to externally assess students up through age 14, but it provides teachers with the option to give students end of stage 3 assessments that are externally graded through the Northern Ireland Council for Curriculum Examinations and Assessments (CCEA). These are largely open-ended assessments that evaluate how students reason, think, and solve problems. CCEA provides multiple assessments for stage 4, according to which pathway a student chooses to follow, including taking the GCSE exam and A-level courses and exams from the U.K. system (whether aiming towards university or a vocational degree) (Council for the Curriculum Examinations and Assessment 2008a, b).

Conclusion

A variety of challenges confront nations seeking to integrate twenty-first century skills into standards, curriculum, assessment, and teaching. An examination of assessment policies and practices in these four nations suggests a range of potential opportunities for evaluating twenty-first century skills in both on-demand tests and curriculum-embedded assessments. The growing move to promote assessment *of, for*, and *as* learning, rather than seeing testing as a separate disjointed element of the education enterprise, may provide opportunities for strengthening the teaching and learning of twenty-first century skills, as well as their assessment.

The growing emphasis on school-based performance assessments in many countries appears to strengthen teaching in which teachers learn more deeply about how to enact standards by participating in scoring and/or reviewing student work. It may also increase curriculum equity, since all students engage in more common activities and instructional supports as part of the required assessments. Some assessment policies also seek to use assessment to strengthen teaching by considering how to provide both feedback and "feedforward" information. They incorporate rich feedback to students, teachers, and schools about what has been learned, and they shape students' future learning by offering opportunities for student and teacher reflection that supports learning to learn. Technology supports for these efforts are becoming increasingly sophisticated and should be shared across states and nations.

Given the critical importance of these initiatives to the teaching and acquisition of twenty-first century skills, the ATC21S project should facilitate countries' efforts to develop optimal policy strategies that integrate school-based assessments of ambitious intellectual performances with large-scale assessments that seek to measure problem-solving, critical thinking, collaboration, and learning to learn in increasingly sophisticated ways.

References

Archer, J. (December 19th, 2006). Wales eliminates National Exams for many students. *Education Week*. Retrieved on September 11th, 2008, from http://www.edweek.org/ew/articles/2006/12/20/16wales.h26.html?qs=Wales.

Buchberger, F., & Buchberger, I. (2004). Problem solving capacity of a teacher education system as a condition of success? An analysis of the "Finnish case. In F. Buchberger & S. Berghammer (Eds.), *Education policy analysis in a comparative perspective* (pp. 222–237). Linz: Trauner.

Chan, J. K., Kennedy, K. J., Yu, F. W., & Fok, P. (2008). Assessment policy in Hong Kong: Implementation issues for new forms of assessment. *The Hong Kong Institute of Education*. Retrieved on September 12th, 2008, from http://www.iaea.info/papers.aspx?id=68

Council for Curriculum Examinations and Assessment. (2008a). *Curriculum, key stage 3, post-primary assessment*. Retrieved on September 12th, 2008, from http://www.ccea.org.uk/

Council for Curriculum Examinations and Assessment. (2008b). *Qualifications*. Retrieved on September 12th, 2008, from http://www.ccea.org.uk/

Dixon, Q. L. (2005). Bilingual Education Policy in Singapore: An analysis of its sociohistorical roots and current academic outcomes. *International Journal of Bilingual Education and Bilingualism, 8*(1), 25–47.

Education Bureau. (2001). Domain on learing and teaching. Hong Kong: Education Department.

European Commission. (2007/2008). Eurybase, The Information Database on Education Systems in Europe, The Education System in Finland.

Finnish National Board of Education. (2007, November 12). *Background for Finnish PISA success*. Retrieved on September 8th, 2008, from http://www.oph.fi/english/SubPage.asp?path=447,65535,77331

Finnish National Board of Education. (2008a, April 30). *Teachers*. Retrieved on September 11th, 2008, from http://www.oph.fi/english/page.asp?path=447,4699,84383.

Finnish National Board of Education. (2008b, June 10). *Basic education*. Retrieved on September 11th, 2008, from http://www.oph.fi/english/page.asp?path=447,4699,4847.

Hautamäki, J., & Kupiainen, S. (2002, May 14). The Finnish Learning to Learn Assessment Project: A concise report with key results. Prepared for the Workshop on Learning-to-Learn Assessment, Brussels. Helsinki: Centre for Educational Assessment, Helsinki University.

Hautamäki, J., Arinen, P., Eronen, S., Hautamäki, A., Kupiainen, S., Lindblom, B., Niemivirta, M., Pakaslahti, L., Rantanen, P., & Scheinin, P. (2002). *Assessing learning-to-learn: A framework*. Helsinki: Centre for Educational Assessment, Helsinki University, and the National Board of Education in Finland.

HKEAA. (2009). School-based Assessment (SBA). Retrieved on August 10th, 2011, from http://www.hkeaa.edu.hk/en/sba

Kaftandjieva, F., & Takala S. (2002). *Relating the Finnish Matriculation Examination English Test Results to the CEF Scales*. Paper presented at Helsinki Seminar on Linking Language Examinations to common European Framework of reference for Languages: Learning, Teaching Assessment.

Kaur, B. (2005). *Assessment of mathematics in Singapore schools—The present and future*. Singapore: National Institute of Education.

Korpela, S. (2004, December). *The Finnish school—A source of skills and well-being: A day at Stromberg Lower Comprehensive School*. Retrieved on September 11th, 2008, from http://virtual.finland.fi/netcomm/news/showarticle.asp?intNWSAID=30625

Laukkanen, R. (2008). Finnish Strategy for High-Level Education for All. In N. C. Soguel, & P. Jaccard (Eds.), *Governance and performance of education systems*. Dordrecht: Springer.

Lavonen, J. (2008). *Reasons behind Finnish Students' Success in the PISA Scientific Literacy Assessment*. University of Helsinki, Finland. Retrieved on September 8th, 2008, from http://www.oph.fi/info/finlandinpisastudies/conference2008/science_results_and_reasons.pdf.

Ng, P. T. (2008). Educational reform in Singapore: from quantity to quality. *Education Research on Policy and Practice, 7*, 5–15.

Qualifications and Curriculum Authority (2009). Assessing pupils' progress: *Assessment at the heart of learning*. Retrieved on May 23, 2009, from http://www.qca.org.uk/libraryAssets/media/12707_Assessing_Pupils_Progress_leaflet_-_web.pdf.

Queensland Government. (2001). *New basics: The why, what, how and when of rich tasks*. Retrieved on September 12th, 2008, from http://education.qld.gov.au/corporate/newbasics/pdfs/richtasksbklet.pdf.

Scottish Qualifications Authority. (2004, March). *Scotland's national qualifications: Quick guide*. Retrieved on September 11th, 2008, from http://www.sqa.org.uk/files_ccc/NQQuickGuide.pdf.

Singapore Examinations and Assessment Board. (2006). *2006 A-Level Examination*. Singapore: Author.

Stage, E. K. (2005, Winter). Why do we need these assessments? *Natural Selection: Journal of the BSCS,* 11–13.

The Finnish Matriculation Examination. (2008). Retrieved on September 8th, 2008, from http://www.ylioppilastutkinto.fi/en/index.html

The Scottish Government. (2008). *Schools: Attainment*. Retrieved on September 11th, 2008, from http://www.scotland.gov.uk/Topics/Education/Schools/curriculum/Attainment.

Victoria Curriculum and Assessment Authority. (2009). Planning for Assessment. http://vels.vcaa.vic.edu.au/support/tla/assess_planning.html.

Welsh Assembly Government. (2008a). *Primary (3–11)*. Retrieved on September 12th, 2008, from http://old.accac.org.uk/eng/content.php?cID=5.

Welsh Assembly Government. (2008b). *Secondary (11–16)*. Retrieved on September 12th, 2008, from http://old.accac.org.uk/eng/content.php?cID=6.

Index

P. Griffin et al. (eds.), *Assessment and Teaching of 21st Century Skills*,
DOI 10.1007/978-94-007-2324-5, © Springer Science+Business Media B.V. 2012