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.

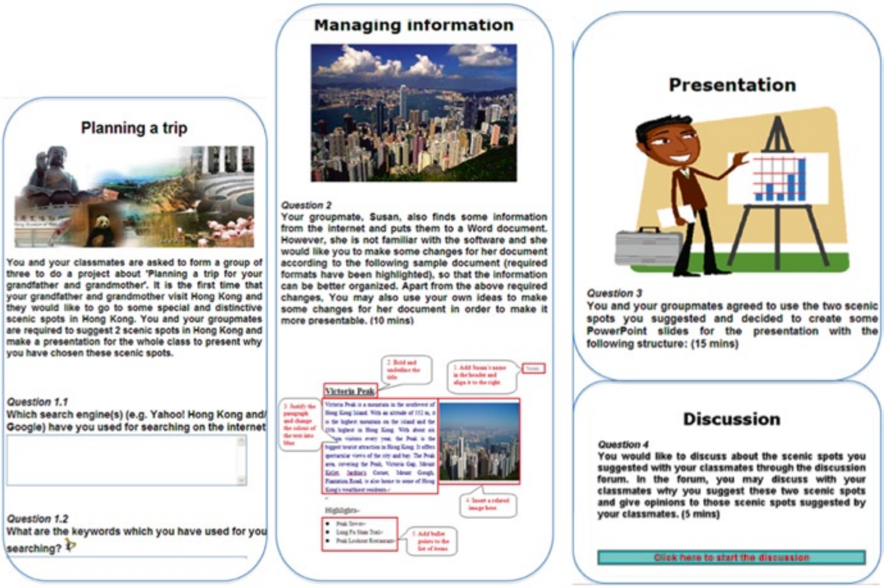
Technology-Based Assessments in Asia

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 8th highest in Hong Kong. With about [illegible] 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)

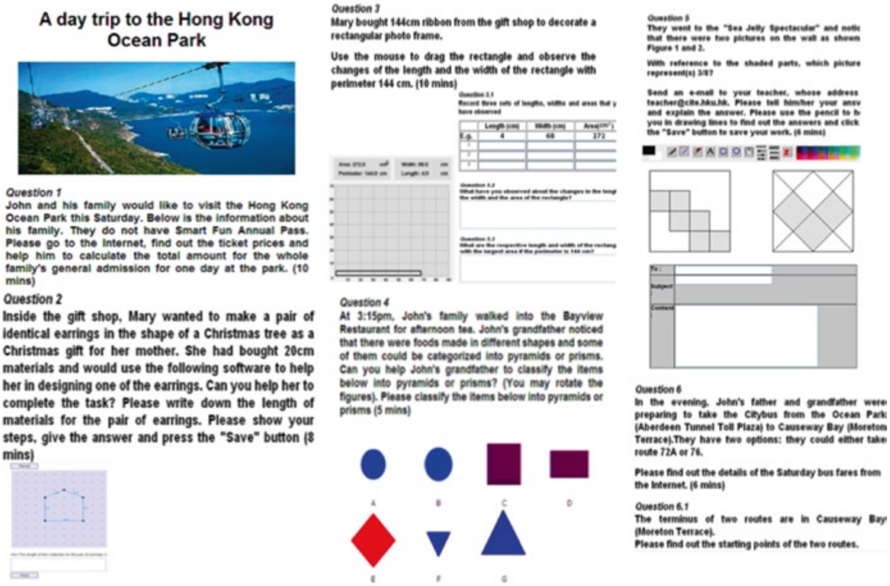
Discussion

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 leng 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?

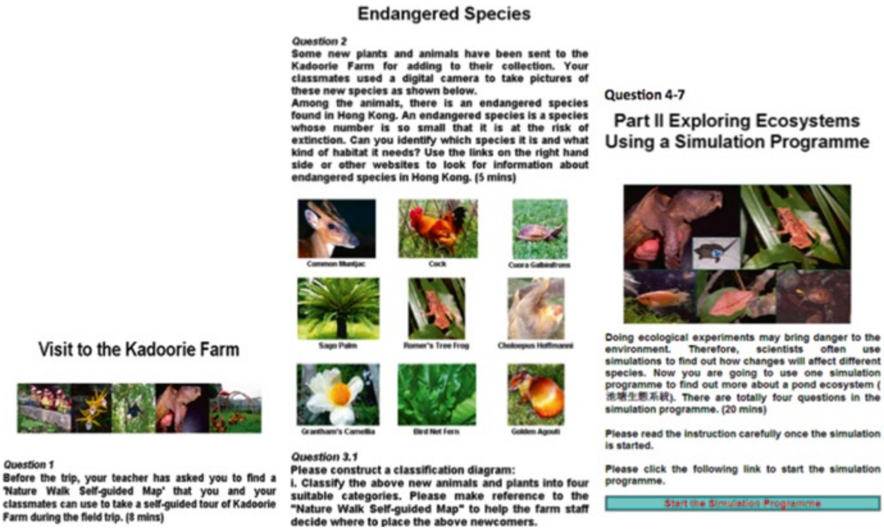
Send an e-mail to your teacher, whose address teacher@cite.hku.hk. Please tell him/her your ansv and explain the answer. Please use the pencil to h you in drawing lines to find out the answers and click the "Save" button to save your work. (6 mins)

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.

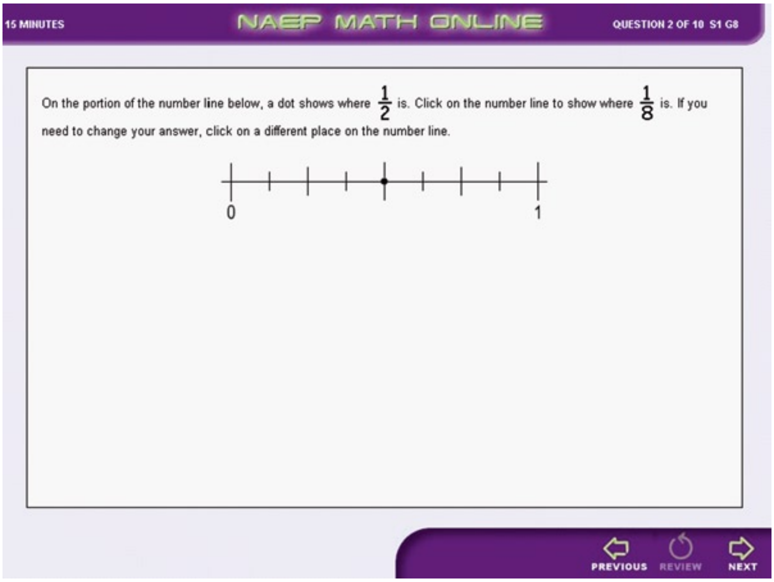
Task Presentation and Response Entry

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)

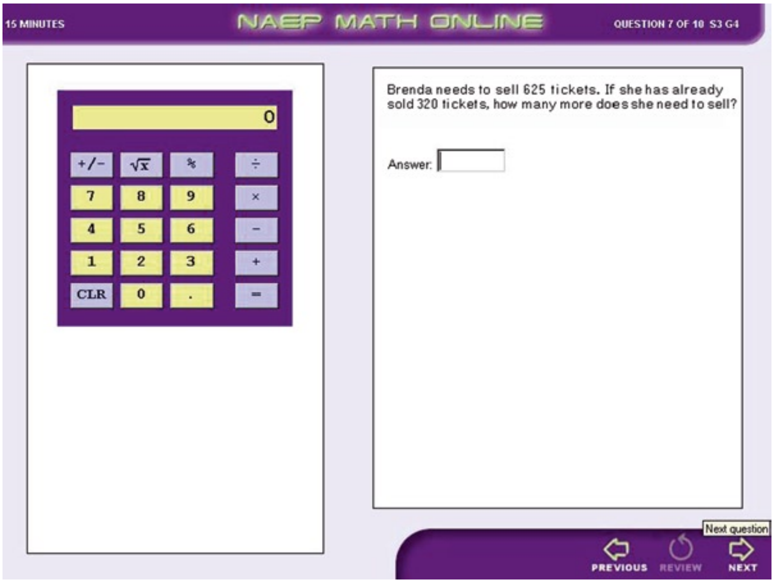
tended to track the manner of problem-solving as it is conventionally practised in classrooms and represented on paper tests, an approach which does not use the computer to maximum advantage. In this case, the computer serves primarily as a task presentation and response collection device, and the key 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:

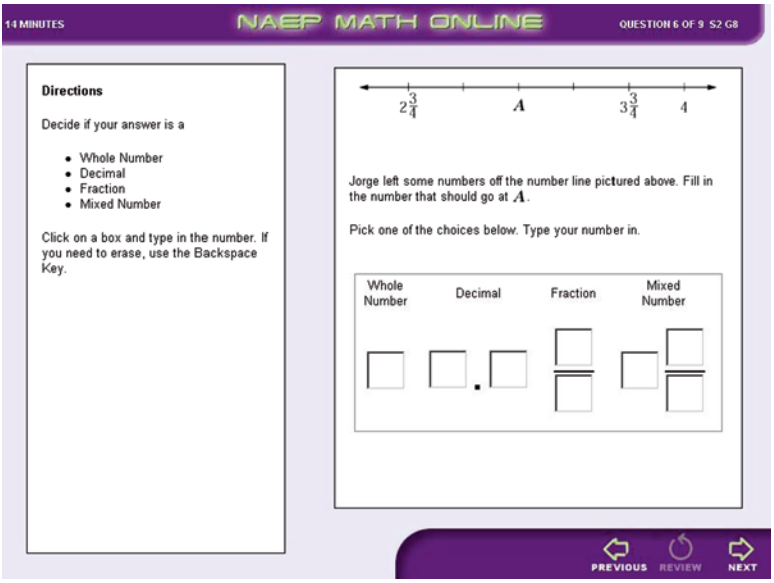
Fig. 4.5 A numeric entry task allowing use of an onscreen calculator (Source: Bennett 2007)

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

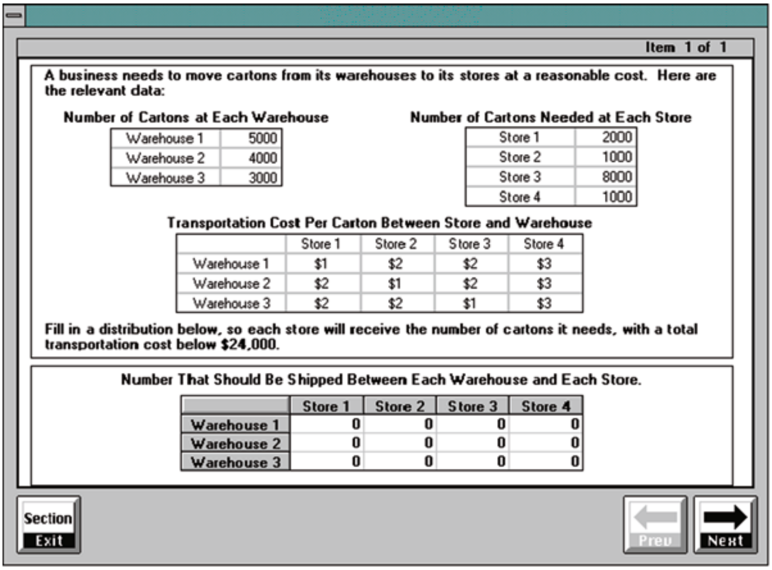
Click on a box and type in the number. If you need to erase, use the Backspace Key.

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

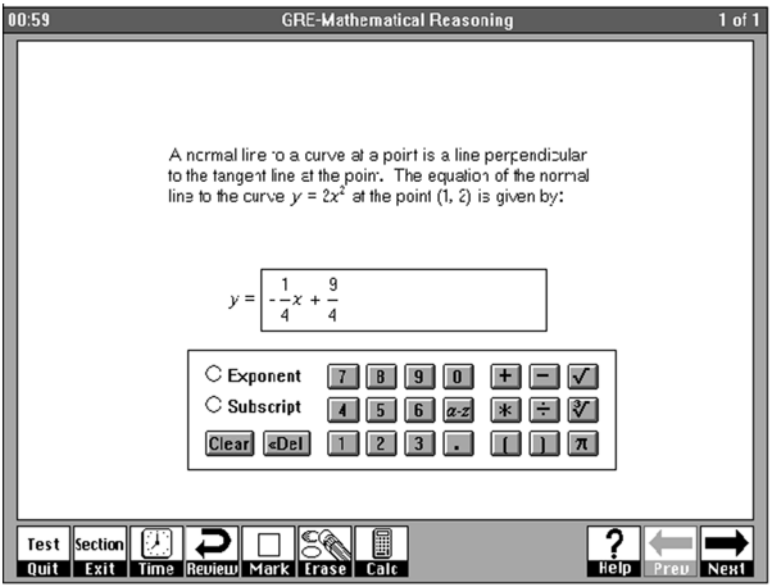
Fig. 4.7 Task with numeric entry and many correct answers to be scored automatically (Source: Bennett et al. (1998). Copyright (c) 1998 ETS. Used by permission)

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.

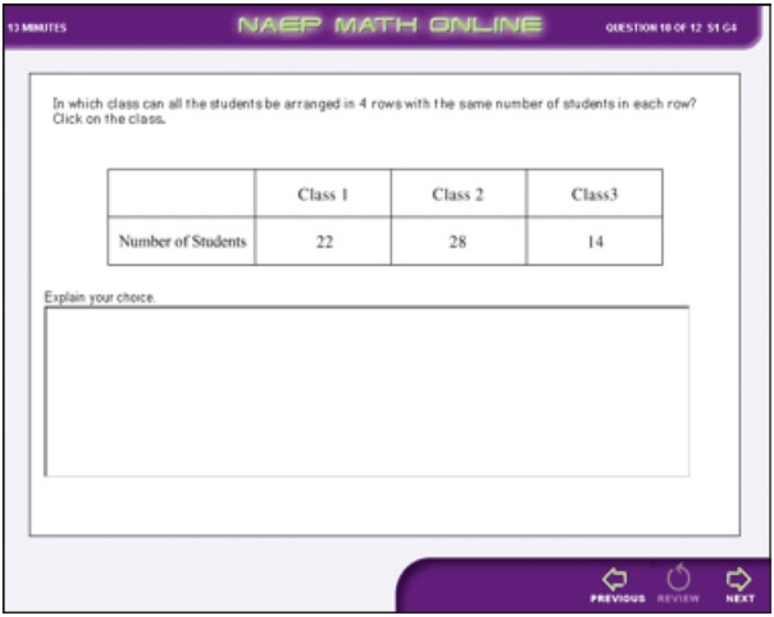
Figure 4.8 shows another type used in graduate admissions research (Bennett et al. 2000). The response type allows questions that have symbolic expressions as answers, allowing, for example, situations presented as text or graphics to be modelled algebraically. To enter an expression, the examinee uses 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.

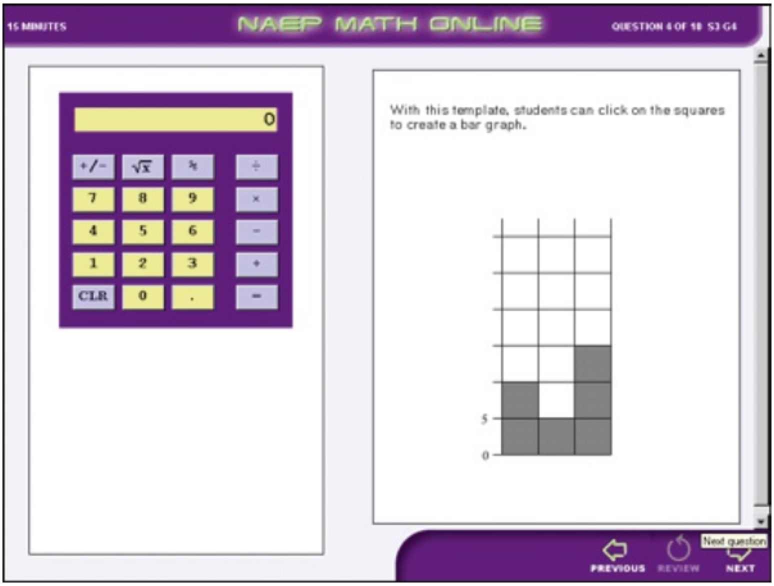
Fig. 4.9 Task requiring forced choice and text justification of choice (Source: Bennett 2007)

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.

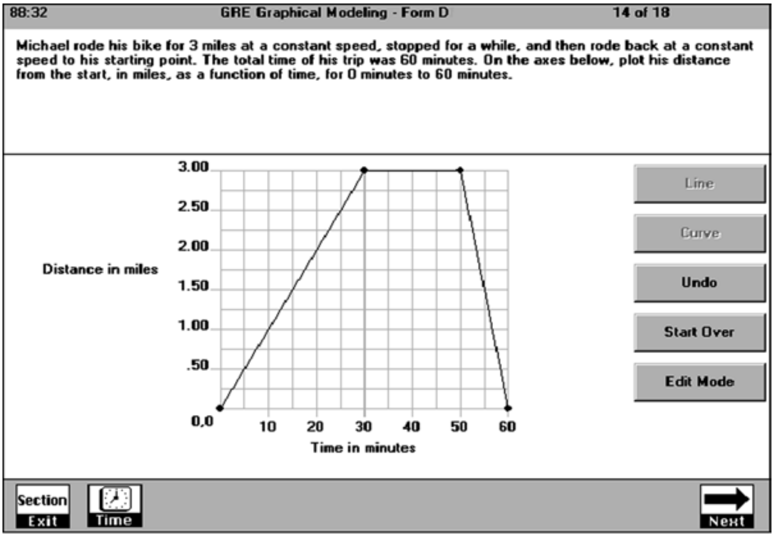
Figure 4.10 shows a NAEP-research format in which the student is given data and then must use the mouse to create a bar graph representing those data. Bars are created by clicking on cells in the grid to shade or unshade a box.

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.

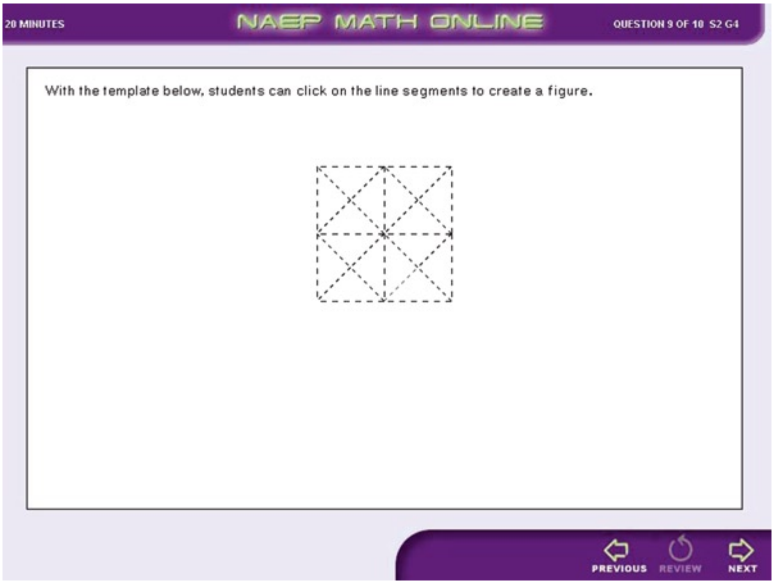
Fig. 4.10 Graph construction with mouse clicks to shade/unshade boxes (Source: Bennett 2007)



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.

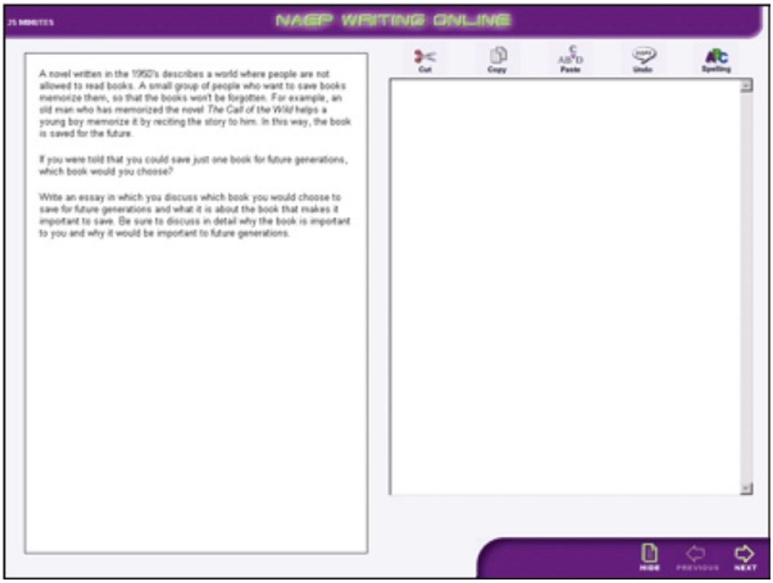
However, there are domains in which technology delivery can make the stimulus dynamic through the use of audio, video or animation, an effect that cannot be achieved in conventional tests unless special equipment is used (e.g. TV monitor with video playback). Listening comprehension is one such domain 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.

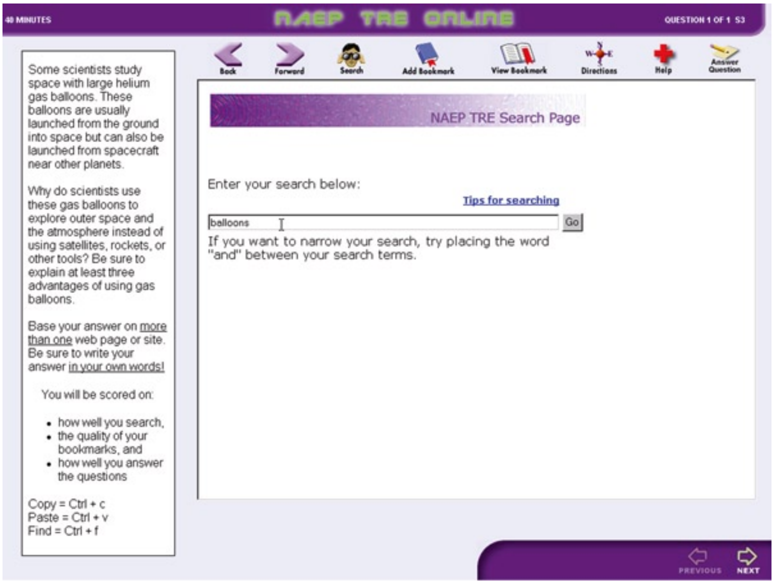
In assessment of writing, as in other domains where a technological tool is employed, a key issue is whether to create a simplified version of the tool for use in the assessment or to use the actual tool. Using the actual tool—in this instance, a particular commercial word processor—typically involves 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.

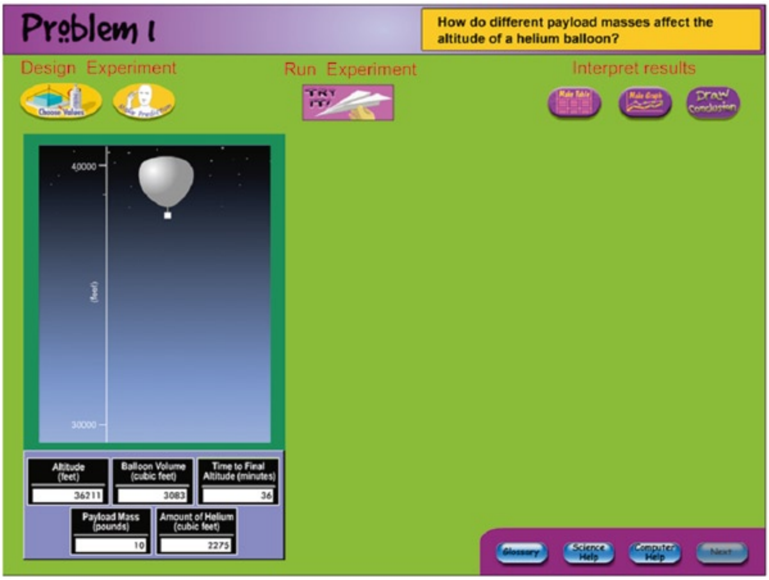
Fig. 4.14 A simulated Internet search problem (Source: Adapted from Bennett and et al. 2007)

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



Problem 1

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.

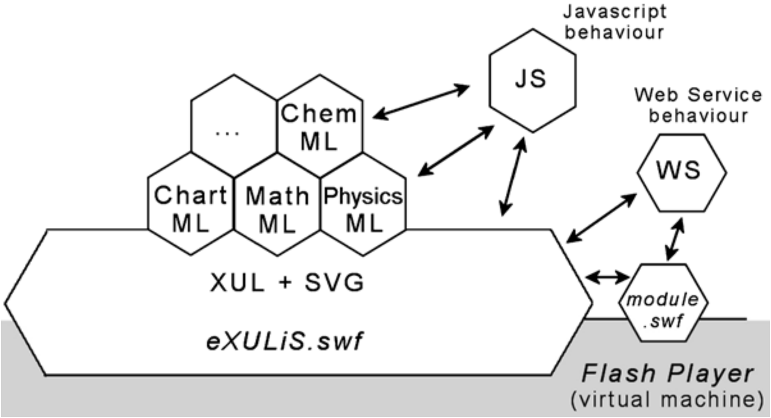


Fig. 4.16 Illustration of eXULiS handling and integrating different media types and services

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 interdisciplinary. 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