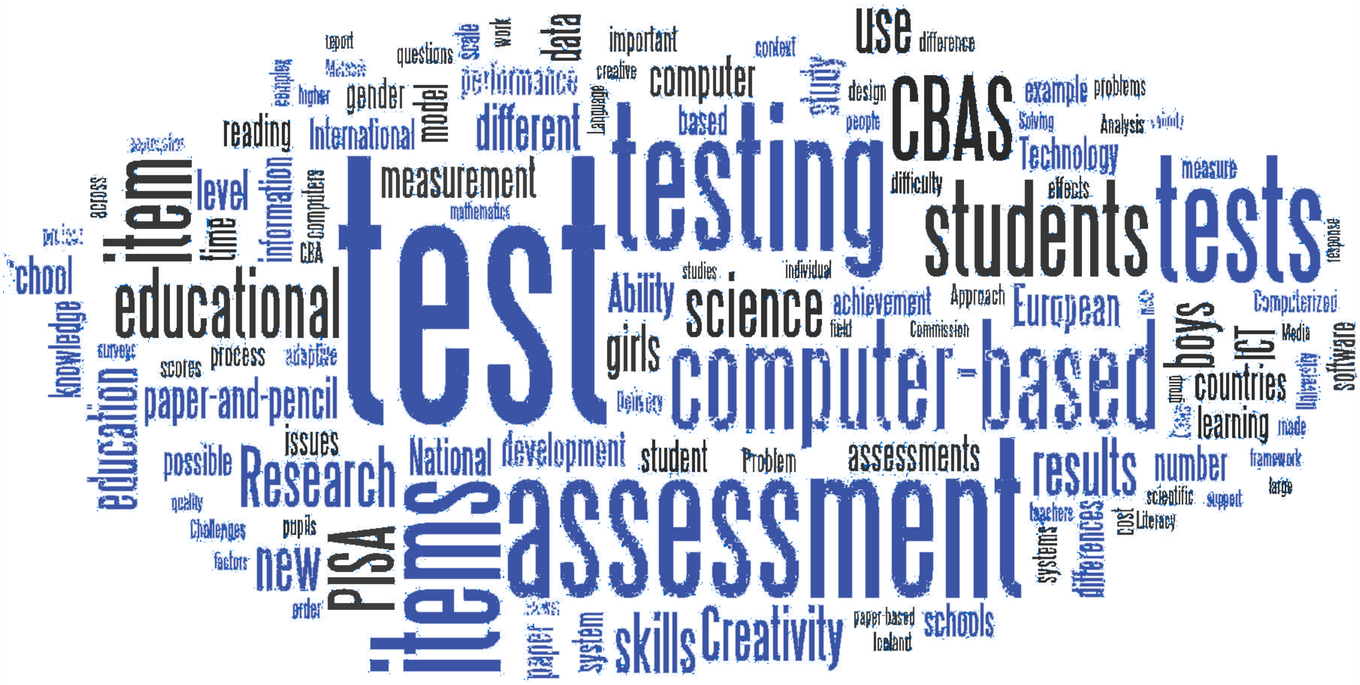
JRC Scientific and Technical Reports

The Transition to Computer-Based Assessment

New Approaches to Skills Assessment and Implications for Large-scale Testing

Friedrich Scheuermann & Julius Björnsson (Eds.)



EUR 23679 EN - 2009

JRC
EUROPEAN COMMISSION

Institute for the Protection and Security of the Citizen

CRELL
Centre for Research on Lifelong Learning

The Institute for the Protection and Security of the Citizen provides research-based, systems-oriented support to EU policies so as to protect the citizen against economic and technological risk. The Institute maintains and develops its expertise and networks in information, communication, space and engineering technologies in support of its mission. The strong cross-fertilisation between its nuclear and non-nuclear activities strengthens the expertise it can bring to the benefit of customers in both domains.

European Commission
Joint Research Centre
Institute for the Protection and Security of the Citizen

Contact information

Address: Unit G09, QSI / CRELL
TP-361, Via Enrico Fermi, 2749; 21027 Ispra (VA), Italy
E-mail: friedrich.scheuermann@jrc.it
Tel.: +39-0332-78.6111
Fax: +39-0332-78.5733

http://ipsc.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

Legal Notice

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

JRC49408

EUR 23679 EN
ISSN 1018-5593
ISBN 978-92-79-11110-5

Luxembourg: Office for Official Publications of the European Communities

Printed in Italy

Foreword

Within the *Lisbon strategy*, Member States agreed to monitor policy implementations with the help of indicators and benchmarks. Regular monitoring allows strengths and weaknesses to be indentified and serve as tools for evidence-based policymaking which is becoming a reality in most European countries. Hence, it is important that the evidence we base our policy on is the best and most accurate possible.

The availability and quality of indicators in the educational field is constantly improving, and more studies and surveys measuring skills have been implemented the last couple of years. International large scale assessments are being realised, not only in Europe but in the whole world. In the PISA survey there were 57 countries that participated in 2006, and 58 countries participated in TIMSS in 2007. In order to ensure good quality of European education we still need to know more about the skills of European citizens. In the future, we therefore expect an increase of surveys covering all age groups from young people to adults.

International surveys in education are expensive. Technology offers new opportunities for innovation in educational assessment, and computers play an important role in order to test efficiently and effectively.

The European Commission has initiated a large scale survey of the general level of competences in reading, listening and writing foreign languages of pupils in the Member States. The Commission wants to make sure that computer-based tests should be made available to all the participating countries and the highest quality of service and open source solutions should be ensured. The references from studies presented in the report will serve as guidelines when the Commission is developing large scale surveys.

Within this given context we welcome the research undertaken in the field of educational measurement focussing on the complex interactions of issues to take into account when making benefit of computer technology from design to the implementation of tests.

The articles of this report highlight the numerous advantages of introducing computers relative to paper-based tests for large-scale testing programs like paperless test distribution and data collection; standardised test administration; permitting more interactive question types and the possibility to create sophisticated tests which include adaptive elements. It has proven to be motivating for students who are given the opportunity to be tested in more realistic settings than paper and pencils can provide. The PISA 2006 cycle included an optional computer-based component assessing scientific competencies. The items developed for the computer-based assessment are based on the same framework as the paper-based assessment. The highly interesting results from participating countries are presented in depth in this report. PISA will seek to deepen the use of computer-based assessments, to allow for a wider range of dynamic and interactive tasks and to explore more efficient ways of carrying out the main tests of student knowledge and skills in reading, mathematics and science.

The work presented here gives examples of experiences with the transition from paper and pencil based tests into the new tools as well as examples of the comparisons of paper and pencil tests and the possible risks to be aware of in this transition. Articles cover the important issues of obstacles and future research needed in the field. Consequently, this report is important in order to reach and implement the new assessment tools.

Several countries are already implementing computer-based tests, and events such as the workshop held in Iceland are extremely important in order to share good practise and learn from each other in this field. The presentations from various countries in Europe and other regions worldwide show the different experiences at country level with the use of computer testing and assessments. These experiences illustrate that there are a lot of complex issues in the transference from paper and pencil tests to using computers. The high level of the contributions in this report is valuable for the activities of the European Commission and other international bodies when developing new surveys as well as for the participating countries who are implementing the surveys.

The report contributes to the increased knowledge base necessary to be developed in the field and emphasise the complexity of this issue and the way forward to develop more effective approaches to computer-based testing and assessments.

Anders Hingel

Head of Unit
European Commission
Directorate-General for Education and Culture
Analysis and Studies Unit

Table of Contents

PART I: ASSESSMENT NEEDS AND EUROPEAN APPROACHES

PART II: GENERAL ISSUES OF COMPUTER-BASED TESTING

PART III: TRANSITION FROM PAPER-AND-PENCIL TO COMPUTER-BASED TESTING

PART IV: METHODOLOGIES OF COMPUTER-BASED TESTING

PART V: THE PISA 2006 COMPUTER-BASED ASSESSMENT OF SCIENCE (CBAS)

Introduction

Technological innovation and new requirements posed by the global economy are affecting the performance of educational systems to a large extent. Societal and structural changes urge educational reforms in many countries where given traditional education does not meet the needs posed to educational institutions and individuals. Advances achieved with integration of educational technology into teaching and learning and new pedagogical approaches enhance the capacities to update to new challenges and it is now up to educational policy to ensure a good match of increasing potentials with skills needed by modern society.

The field of skills assessment has therefore been gaining increasing awareness for a number of years in international research and practice for various reasons:

(1) Policy-makers are asking for accurate measures of the current level of their citizens' competences and for ways of monitoring changes.

(2) Companies want to know about the skills of their employers for staff development purposes or hiring needs.

(3) In educational practice there is more emphasis given to assessment. With increasing availability of information and communication technologies (ICT), new possibilities are provided to assess learning processes and outcomes which are more effective than was possible with traditional assessment/testing instruments.

(4) Finally, it is the learner him/herself who is now better able to monitor his state of learning and progress. There is, furthermore, a need to provide him/her with testing means based on availability, habits and preferences.

At a general level, the question arises to what extent ICT can contribute to support assessment activities in these given contexts and what policy can do in order to ensure effective implementation. It is both, assessment "of" and assessment "for" learning which is gaining more interest with the availability of ICT.

International surveys and computer-based testing

Future international surveys are going to introduce new ways of assessing student achievements. Electronic tests, especially adaptive ones can be calibrated to the specific competence level of each student and become more stimulating, going much further than can be achieved with linear tests made up of traditional multiple choice questions. Simulations also provide better means of contextualising skills to real life situations and provide a more complete picture of the actual competence to be assessed. However, a variety of challenges require more research into the barriers posed by the use of technologies, e.g. in terms of computer requirements, performance and security.

As far as policy and monitoring tools are concerned there is an increasing trend to carry out comparative surveys both at a national and international level. The OECD PISA (Programme for International Student Assessment) is the most prominent study of this type in education. It is running now in more than 60 countries and has been going on since the year 2000. PISA is in many ways different from former international surveys, such as the TIMSS, which is curriculum based, in that it attempts to assess the skills and competencies each learner needs for further study and for success in the future. Although the basic skills assessed are similar (reading, math, science) their definition is different and broader and the tasks are put into the context of everyday situations which most people will at one time or another have to deal with.

PISA is therefore more focused on the new skills and competencies needed in a rapidly changing world and this focus requires ongoing changes and adaptations of the study to meet the changing needs of the educational systems. An important part of the study is therefore exploring and trying out new ways of measuring educational outcomes

PISA offers a good opportunity to take a closer look at the challenges posed for international comparative surveys and to identify the critical areas for further work needed for the following reasons:

- PISA has already started the process to implement computer-based modules (electronic reading) for future tests. All in all, it is intended to move completely to Computer-based Testing (CBT) by the year 2015 and there are still a lot of open questions to be discussed.

- Field studies and the 2006 CBAS study offer a unique opportunity to look at results of CBT in comparison to the Paper-and Pencil (P&P) version and to reflect about interpretations and consequences

- Overall, there is a challenge in how to move to CBT but keeping the visibility of trends from previous P&P tests.

- In some ways the two goals of preserving trends and going to CBT based methods are incompatible. Changing the nature of tests generally ruins trends and keeping tests unchanged is unsatisfactory as the definition of the skills and competencies that should be measured are continually changing and evolving. This paradoxical situation has to be overcome and in order to do so both new technologies and new conceptualizations of what is being measured have to be explored.

Despite the benefits of computer-based testing European countries are currently facing the challenge to shift from traditional testing to computer-based assessment approaches and to organize a smooth process of transitioning. Due to the given complexity of issues to take into account there is no simple and clear solution on the right approach and the right methodologies for computer-based testing. At the end, it is a matter of finding a compromise solution in order to combine potentials and constraints for different areas, such as technological, economic and measurement considerations. The amount of human effort and costs are directly related to test design, needing to be carefully thought about and to be related to expected gains for skills assessment. In financial terms, required budgets and country contributions for carrying out the survey have to be low as more surveys have to be delivered both at country and European/international level in general.

The political interest in eAssessment research

In 2006 the European Parliament and the Council of Europe have passed recommendations on key competences for lifelong learning and the use of a common reference tool to observe and promote progress in terms of the achievement of goals formulated in “Lisbon strategy” in March 2000 (revised in 2006, see http://ec.europa.eu/growthandjobs/) and its follow-up declarations in selected areas (see Bjerkestrand, this volume). For those areas which are not already covered by existing measurements effective instruments are now needed for carrying out large-scale assessments in Europe. In this context it is hoped that electronic testing could improve the effectiveness of the needed assessments, i.e. improve identification of skills, and their efficiency, by reducing costs of the whole operation.

The Joint Research Centre of the European Commission in Ispra, Italy, (http://crell.jrc.ec.europa.eu) is supporting the DG Education and Culture in the preparation of future surveys. Currently, the Centre for Research on Lifelong Learning (CRELL) is carrying out a research project on modes and platforms for computer-based testing and to analyse effectiveness of software implementations for large-scale surveys. The research approach is framed by the need to assess skills of population groups in Europe at a large scale and to achieve accurate and comparable results for further benchmarking. Therefore, emphasis is given to tools for summative assessment and objective measurement as the basis of research activities on e-assessment.

The Educational Testing Institute, in Reykjavik, Iceland is in charge of carrying out national school assessment and has had on its agenda now for some time the transition of conventional paper and pencil based tests over to computerized adaptive testing. The Institute is also responsible for carrying out the Icelandic parts of a number of international surveys, among others the PISA study and that experience has further strengthened the CBT plans. This plan for going over to computer based testing is partly a response to requests from the school system where more and more emphasis is being put on individualised learning and teaching and an adaptive electronic approach to testing, fits that approach very well. There have been in recent years growing concerns about the conventional national paper and pencil tests and all stakeholders are presently very enthusiastic about utilizing the new methods of electronic testing.

Within the scope of these national assessment activities and the intention to move to computer-based assessment in a short period of time a common research project was formulated aiming at monitoring the transition phase in selected subject areas. The national activities in Iceland and recent discussions on how to proceed in this transition process offered a unique opportunity for a case study which would demonstrate significant insights for the benefit of other countries facing a similar situation.

The following research questions were formulated:

- What is the added value of adaptive testing to traditional computer-based methods?
- To what extent does CBA improve methods for skills assessment?
- What are the challenges for e-Testing platforms, when actually applied in large-scale surveys?
- What are the challenges identified, e.g. when internet-based delivery modes are applied?
- Which quality aspects for CBA deployment apply in order to define quality standards for electronic assessments?

Expert workshop in Reykjavik

Within this context of continuous reflections and in the light of the complexity of issues to be looked at a research workshop was organized by the Educational Testing Institute and CRELL on the “The Transition to Computer-Based Assessment - Lessons learned from the PISA 2006 Computer Based Assessment of Science (CBAS) and implications for large scale testing”. The event was held in Reykjavik, Iceland from September 29th to 1st October 2008. Over 100 international high-level experts from research and practice of educational measurement attended the workshop in order to discuss relevant experiences and research along the workshop themes in the context of both the results presented on PISA 2006 Computer-Based Assessment of Science (CBAS) and the intention of Iceland to shift from traditional paper-and-pencil testing to computer-based assessment. The event was carried out in cooperation with the Indicators and Analysis Division of the OECD which is in charge of PISA implementation.

The main themes of the workshop were oriented towards some of the most important assessment aspects when moving from traditional to computer-based testing, such as the comparison between paper and pencil tests and computer based assessment (methodologies, approaches, effects etc.), gender differences (use of media and differences in results) and adaptive vs. linear computer-based assessment in the context of both the results presented on PISA 2006 Computer-Based Assessment of Science and the intention of Iceland to shift from traditional paper-and-pencil testing to computer-based assessment.

During the 1st day most discussions were focused on specific issues of electronic testing, such as computerised and adaptive testing in educational assessment, experiences from large scale computer based testing in the USA and European approaches to introduce computer-based testing at a national level. The new Danish National Test is an interesting example of explaining the reasons and potential benefits in the context of challenges identified with its introduction (next year). Apart from experiences presented from other European countries (e.g. Germany, Norway, Croatia, Hungary), other talks were held on delivery platforms, economic models and comparison of results of Paper-Pencil vs. computer-based tests. Furthermore, emphasis was also given to European approaches and future intentions as coordinated by the European Commission (Coherent Framework of Indicators and Benchmarks and implications for CBA; computer-based testing of foreign language skills, and computer-based measurement of creativity).

The second day was mainly devoted to discussions about the PISA 2006 CBAS pilot study. Since the results were presented for the first time there was a high interest in looking into the outcomes of the test which was finally carried out in only 3 countries (Denmark, Iceland, Korea). Overall, no major difference was encountered in terms of outcomes on the paper and pencil test and computer-based test. However, boys outperformed girls on the computer-based assessment of science in all countries but performance cannot easily be linked to motivation, enjoyment or familiarity with computers.

In addition to these results an outline was given on the plans for PISA 2009, and ideas for 2012/2015. One of the major challenges will be how to ensure comparability of PISA results without losing a key feature to observe trends over the years of PISA assessments.
Further discussions on the implementation of computer-based assessment in Iceland were held during the 3rd day of the meeting. Apart from given introductions to the Icelandic system possible approaches were discussed as well as some further consideration was given to software requirements and available platforms.
Further information on the event and presentations held can be found on the web-site of the event at: http://crell.jrc.eu.europa.eu.

About this report

This report represents a combination of paper presentations from the conference with conclusions on the basis of discussions held, and a variety of additional articles in relevant areas where further information was requested. The document is clustered in 5 thematic groups: the first set of articles is dealing with assessment needs and European approaches as far as comparative surveys on the use for educational policy are concerned. The next group of articles looks into more general issues relating to computer-based testing, experiences and challenges posed. Then, a third section is dedicated to experiences and reflections on the transition from paper-and-pencil to computer-based testing. Computer-based methodologies of test design, testing and interpretation of results are discussed in section four, while the last part is focusing on the results of the PISA study on Computer-based Assessment of Science (CBAS).

Section 1: Assessment needs and European approaches

Robert Kozma provides an interesting contribution on the changing context of skills and their measurement and needs for future assessments. The European Coherent Framework of Indicators and Benchmarks is then introduced by *Oyvind Bjerkestrand*. This framework constitutes the overall scope of European approaches in skills identification and their measurement in the context of the "Lisbon agenda" which are currently in the process of being undertaken or intended during the next years. The results are yearly published in the Progress Report of the European Commission which serves as the reference point for monitoring the educational change in Europe (http://ec.europa.eu/education/lifelong-learning-policy/doc34_en.htm). *Ernesto Villalba* then tries to look into the implications of computer-based measurement in the area of "creativity in education" which constitutes an ill-defined area where, however, it is hoped that measurements can contribute to raise awareness and stimulate improvements of creative processes in education. From his view CBT could provide better answers to the challenges posed when assessing such processes as it could be expected from Paper-Pencil testing.

Section 2: General issues of computer-based testing

The second section starts with a general overview on CBT by *Brent Bridgeman*. Advantages and challenges are also contrasted to traditional methods such as paper-pencil testing. *Jakob Wandall* reports about experiences made in Denmark with introducing computer-based tests and especially, computer-adaptive testing. *Elli Moe* reports about experiences made relating to the introduction of computerized tests for languages (English) in Norwegian schools. Although, overall, these tests were positively accepted, there are still challenges to in terms of productive skills to be included in the scope of assessment. This was also recognized for one of the future surveys to be carried out at a European level on language skills of school children where productive skills will not be considered in the first assessment round. The computerized approach of this survey is presented by *Jostein Ryssevik* who presents the technical concept of the implementation. A further technology-oriented view is provided by *Sam Haldane* who describes delivery platforms used for the PISA CBAS pilot study and a national assessment project in Australia, including given intentions for the PISA 2009 Electronic Reading Assessment (ERA). Apart from implementation and economical aspects he also points out the challenge of ensuring security and the requirements posed. Inappropriate infrastructure in school environments is still a challenge to be met with

Internet-based test delivery. Finally, the section is complemented by a view of *Klaus Reich* and *Christian Petter* on accessibility issues and design considerations taking into account the needs of user groups with special needs, as presented in existing guidelines and standards,

Section 3: Moving from Paper-and-Pencil Testing to Computer-based Testing

Implications related to the transition from paper-and-pencil to computer-based testing are focused in this section. First *Gerben van Lent* provides an overview on risks and benefits when deciding to move to CBT. He takes a closer look at the decision-making processes and offers a model for dealing with the complexity of issues to take into account. Transformative assessment as the approach to meet the challenges and needs of 21st century learners is introduced by *Martin Ripley.* He presents examples which demonstrate innovative and promising approaches to the use of computer technology in assessment which help to get an orientation about the potentials in contrast to traditional non-transformational assessment approaches. The arguments made are supported by *Katerina Kikis-Papadakis* who worries about a wrong understanding of eAssessment as migration from paper-and-pencil testing to CBT leading to disadvantages of certain learner groups. As a first and most important step she urges a dialogue needed among all stakeholders in school education about what is to be achieved with the use of ICT for learning supported by the research community in the implementation of innovative practices in education (and assessment) which also need to be reflected in radical reforms of the curricula. In similar ways, *René Meijer* points out the negative consequences of “substitution” strategies which miss the point to make real benefit of the potentials given by computer technologies. He suggests to clearly identify the assessment purpose in the context of stakeholders and proposes authenticity, transparency and multiplicity as general principles to take into account. A economical perspective on the decision of moving to computer-based testing is offered by *Matthieu Farcot* and *Thibaud Latour.* They present a general framework for the analysis of costs and relate these to 4 different scenarios: paper-pencil, computer-aided, computer-based with taylor-made system and computer-based on the basis of a general platform. *Vesna Buško* then reports about the experiences made with moving from paper-pencil testing to computerized tests. She reflects about the implications from a Croatian perspective where the introduction of CBT is not yet regarded as a realistic scenario for the moment. Finally, *Benő Csapó, Gyöngyvér Molnár* and *Krisztina R. Tóth* take a comparative look into paper-and-pencil and online assessment of reasoning skills and first results of a pilot study carried out in public education in Hungary.

Section 4: Methodologies of Computer-based Testing

This section looks into the methodologies, constructs and media equivalence issues. *Nathan A. Thompson & David J. Weiss* first introduce to the various approaches of computer-based, i.e. computer-adaptive testing. Advantages and disadvantages of test delivery modes are discussed from the perspective of what is possible today and tomorrow with specific view on technologies and infrastructures. A case study of computer-adaptive testing as entry exam for a primary school teacher training college in the Netherlands is the presented by Theo J.H.M. Eggen & Gerard J.J.M. Straetmans. They argue for solutions combining the concept modern testing theory (IRT) and computer-controlled testing and present successful examples of good practice with applying computer-adaptive testing when sufficient time and resources are made available for preparation. Measurement devices and methods are then further explored by *Oliver Wilhelm* who focuses on selected validity fallacies that would deserve more attention in psychometric research. *Patrick Kyllonen* reports about advances made in automated item generation and discusses progress made relating to increased efficiency and convenience, and the assessment of new constructs (such as creativity, leadership etc.) using new methods which are not easily implemented with paper-and-pencil technology. Complex problem solving is an example which is investigated and presented more in detail by *Samuel Greiff* and *Joachim Funke.* They introduce to a testing instrument for dynamic problem solving, covering cognitive facets that could not be tested yet with traditional testing of cognitive abilities. Finally, the issue of equivalence of tests across media is discussed by *Ulrich Schroeders.* This topic is highly relevant when different devices are used and compared by test results as undertaken in PISA 2006 vs. PISA 2006 CBAS and still represents a crucial aspect of further research needed on the question if that what is measured actually represents that what was intended to be assessed.

Section 5: The PISA Computer-based Assessment of Scientific Literacy

The last chapter is dedicated to experiences made with the PISA 2006 pilot study in computer-based assessment of science (CBAS) which, after a series of preliminary field trials, constitutes the first study on the use of computer technology for testing in the context of comparative large-scale surveys and PISA. On the basis of CBAS results *Ron Martin* first explores the nature of computer-based testing in comparison with traditional paper-and-pencil testing in the subject area concerned. He discusses the reasons why the 2006 CBAS outcomes could not be scaled with the paper-and-pen outcomes and describes the (computer) culture-related limitations in international comparisons. An extensive review of CBAS results is then provided by *Almar M. Halldórsson, Pippa McKelvie & Júlíus K. Björnsson* that provide a closer look on the interaction between gender, test modality and test performance among participating countries. There are a variety of possible explanations why Icelandic boys behaved better in computerized tests but there is still no clear interpretation possible and no final conclusion can be drawn yet from these results. CBAS experiences from Korea are reported by *Mee-Kyeong Lee.* She reports about advantages of the CBAS implementation and challenges identified. She also confirms earlier statements about the need of being cautious about a variety of aspects related to the transition from the paper-and-pencil test to computer-based testing. Finally, *Helene Sørensen & Annemarie Møller Andersen* discuss about the CBAS test in Denmark and a significant gender difference observed which clearly indicated a better performance of male participants.

Lessons to be learnt

The articles demonstrate that there is still a lot of research needed in order to ensure a convincing approach to computer-based testing in future surveys. Although there is a clear trend towards computer-based testing for a variety of reasons mentioned earlier, it can be concluded that the complexity of inter-related issues to take into account increases with the use of computer technology. It would be a wrong conclusion to draw a negative picture on the real benefits of computer-based testing and therefore, it would be a wrong consequence to stop further activities due to the complexity of challenges to overcome and variety of problems to solve. The potential benefits far outweigh the problems.

Several dimensions have been covered by the workshop. The coverage of the workshop was important because it allowed exploring the overall situation of computer-based assessment in Europe and elsewhere in the world in order to derive relevant quality criteria, the challenges posed by Computer Based Assessment (CBA), as well as important future research topics. This has hopefully helped to identify the barriers posed and further needs with respect to CBA.

More and more of the active working hours of most people are spent using computer technology and the successful resolution of an increasing number of the problems and tasks people have to solve in modern society is dependent on this. However the educational systems in most developed countries are not following this trend to the same extent, while computer and internet use reaches unprecedented heights in companies and homes. A number of surveys, the PISA among others, show that over the last decade the use of information technology has been at a standstill in schools, that traditional methods appear to be dominant there.

One way of changing this is to move testing in schools over to electronic media, thereby enriching the testing experience and making the test results more useful for teachers and students. Better and more focused tests with more relevant results are surely the way to go, tests which are more relevant for the needs of the future and which can be adapted to the rapidly changing needs and skills the future is going to require. If the current volume can be a small step in that direction then the effort has been worthwhile.

Transforming Education: Assessing and Teaching 21st Century Skills

Assessment Call to Action

Robert Kozma
Intel, Microsoft, and Cisco Education Taskforce

Purpose of this paper

The structure of global economy today looks very different than it did at the beginning of the 20th century, due in large part to advances in information and communications technologies (ICT). The economy of leading countries is now based more on the manufacture and delivery of information products and services than on the manufacture of material goods. Even many aspects of the manufacturing of material goods are strongly dependent on innovative uses of technologies. The start of the 21st century also has witnessed significant social trends in which people access, use, and create information and knowledge very differently than they did in previous decades, again due in many ways to the ubiquitous availability of ICT.

These trends have significant implications for education. Yet most educational systems operate much as they did at the beginning of the 20th century and ICT use is far from ubiquitous. Significant reform is needed in education, world-wide, to respond to and shape global trends in support of both economic and social development. What is learned, how it is taught, and how schools are organized must be transformed to respond to the social and economic needs of students and society as we face the challenges of the 21st century. Systemic education reform is needed that includes curriculum, pedagogy, teacher training, and school organization.

Reform is particularly needed in education assessment—how it is that education and society more generally measure the competencies and skills that are needed for productive, creative workers and citizens. Existing models of assessment typically fail to measure the skills, knowledge, attitudes and characteristics of self-directed and collaborative learning that are increasingly important for our global economy and fast changing world. New assessments are required that measure these skills and provide information needed by students, teachers, parents, administrators, and policymakers to improve learning and support systemic education reform. To measure these skills and provide the needed information, assessments should engage students in the use of technological tools and digital resources and the application of a deep understanding of subject knowledge to solve complex, real world tasks and create new ideas, content, and knowledge.

Efforts to transform assessments have been hindered by a number of methodological and technological factors and these barriers must be addressed. In issuing this call to action to political, education, and business leaders, Cisco, Intel, and Microsoft argue for an international multi-stakeholder project that will:

- Mobilize the international educational, political, and business communities around the need and opportunity to transform educational assessment—and hence, instructional practice—and make doing so a global priority.
- Specify high-priority skills, competencies, and types of understanding that are needed to be productive and creative workers and citizens of the 21st century and turn these specifications into measurable standards and an assessment framework.
- Examine innovative ICT-enabled, classroom-based learning environments and formative assessments that address 21st century skills and draw implications for ICT-based international and national summative assessments and for reformed classroom practices aligned with assessment reform.
- Identify methodological and technological barriers to ICT-based assessment, support the specification of breakthrough solutions that are needed to measure 21st century skills, and derive implications for the scaling up of ICT-enabled classroom learning environments.
- Support the implementation of these standards and breakthrough methodologies, pilot test them in selected countries, and make recommendations for broader educational assessment reform.

This paper presents the rationale for such a project, reviews the current state of art in the assessment of 21st century skills, and identifies the current barriers and problems in developing transformational 21st century assessments. It also provides an action plan by which multiple stakeholders can work together, identify problems, share knowledge, build on current efforts, and create breakthrough solutions to reform assessment and transform education.

Project rationale

Major Changes in the Economy and Work

Restructured economy. Over the past four decades, there have been dramatic shifts in the global economy. One shift has been from the manufacture of goods to provision of services. Research at the UCLA Anderson School of Management documents this shift (Kamarkar & Apte, 2007; Apte, Kamarkar & Nath, in press). In every country of the world's 25 largest economies, services either account for more than 50% of the GNP or they are the largest sector in the economy. But a more significant shift has been from an economy based on material goods and services to one based on information and knowledge. For example in the U.S., the production of material goods (such as automobiles, chemicals, and industrial equipment) and delivery of material services (such as transportation, construction, retailing) accounted for nearly 54 % of the country's economic output in 1967. By 1997, the production of information products (such as computers, books, televisions, software) and the provision of information services (financial services, broadcast services, education) accounted for 63% of the country's output. Information services alone grew from 36% to 56% of the economy during that period.

Restructured work. The structure of companies and the nature of work have also changed. Organizational structures have become flatter, decision making has become decentralized, information is widely shared, workers form project teams, even across organizations, and work arrangements are flexible. These shifts are often associated with increased productivity and innovativeness. For example, a U.S. Census Bureau study (Black and Lynch, 2003) found significant firm-level productivity increases that were associated with changes in business practices that included reengineering, regular employee meetings, the use of self-managed teams, up-skilling of employees and the use of computers by front-line workers. A U.S. Department of Labor study (Zohgi, Mohr, & & Meyer, 2007) found a strong positive relationship between both information sharing and decentralized decision making and a company's innovativeness. Yet typical instructional practices in schools do not include collaboration, information sharing, or self-management.

Enabled by ICT. These changes in organizational structures and practices have been enabled by the application of ICT for communication, information sharing, and simulation of business processes. Recent studies of firms (Pilat, 2004; Gera & Gu, 2004) found significant productivity gains associated with specific ways that technology is being used. The greatest benefits to a firm are realized when ICT investments are accompanied by other organizational changes, such as new strategies, new business processes and practices, and new organizational structures. Yet ICT use in schools is most often incidental and supplements traditional practices and organizational structures rather than new strategies and structures.

Require new skills. These changes in organizational structure and business practices have resulted in corresponding changes in hiring practices of companies and the skills needed by workers. A Massachusetts Institute of Technology study (Autor, Levy, & Murnane, 2003) of labor tasks in the workplace found that commencing in the 1970's, routine cognitive and manual tasks in the U.S. economy declined and non-routine analytic and interactive tasks rose. This finding was particularly pronounced for rapidly computerizing industries. The study found that as ICT is taken up by a firm, computers substitute for workers who perform routine physical and cognitive tasks but they complement workers who perform non-routine problem solving tasks. Because repetitive, predictable tasks are readily automated, computerization of the workplace has raised demand for problem-solving and communications tasks such as responding to discrepancies, improving production processes and coordinating and managing the activities of others. The net effect is that companies in the U.S. and other developed countries (Lisbon Council, 2007) are hiring workers with a higher skill set. In the 21st century economy and society, the memorization of facts and implementation of simple procedures is less important; the ability to respond flexibly to complex problems, to communicate effectively, to manage information, to work in teams, to use technology, and to produce new knowledge is crucial. These capabilities are rarely taught in schools or measured on typical assessments.

Major Changes in Society and Everyday Life

Widespread access to ICT. Access to ICT is spreading widely across the world and affecting the everyday lives of people. According to 2005 World Bank figures, a majority of households in most of the world's largest economies have immediate access to television, cell phones, and the internet. Yet ICT availability in most schools is limited and often ICT is kept in closets or dedicated laboratories.

New patterns of information use. The pervasiveness of ICT has changed the way people access information and other people, as well as the way they use information and create new knowledge. People use the internet to find jobs, look for mates, stay in touch with relatives, do their shopping, book flights, run for office, solicit donations, share photos, post videos, and maintain blogs. Studies in North America, Europe, and Asia document that large numbers of people use the internet regularly and do so to conduct online purchases, use online chat or messaging and download music or movies, play games, exchange email, conducting banking transactions, and searching for information. In the U.S., according to the Pew Internet and American Life Project, more than half of all Americans turn to the internet to find answers to common problems about health, taxes, job training, government services (Fallows, 2008). And more and more Americans are using the internet to access multimedia material and to create digital content (Rainie, 2008; Lenhart, Madden, Macgill, & Smith, 2007). In the U.K., 49% of the children between the ages of 8-17 who use computers have an online profile; 59% use social networks to make new friends (Ofcom, 2008). Students come into classrooms with new ICT skills and competencies but they are rarely drawn on in the formal curriculum nor are students able to use these skills to collaboratively solve complex, real world problems.

Little Change in Education

Businesses, entire economies, and society generally have made dramatic changes over the past decades, much of it enabled by the wide-spread use of ICT. But education systems have been slow to respond. For the most part, curricula, pedagogy, school organization, and assessment are much like they were at the turn of the 20th century. While people outside of school work flexibly in teams, use a variety of digital tools and resources to solve problems and create new ideas and products, students in schools meet in structured classrooms at specified times; teachers cover the standard content by lecturing in front of the class while students listen; students work individually and reproduce this knowledge on assessments; and their use of ICT is limited. This pattern is global. A recent international survey of teachers in 23 countries in North America, Europe, Asia, Latin America, and Africa (Law, Pelgrum, & Plomp, 2008) found that the three most common pedagogical practices were having students fill out worksheets, work at the same pace and sequence, and answer tests. ICT was rarely used and the applications used most often were general office software, followed by tutorial or drill and practice software.

At the same time, there are new models of technology-rich learning environments and formative assessments that engage students in collaborative problem solving and the production of creative works. Yet the use of these new models is still rare, in part because traditional assessments are inadequate to measure the outcomes of their application.

The Need to Transform Assessment

Current assessments reflect typical pedagogical and assessment practices found in classrooms but they are also a key determiner of what students learn in classrooms and how that is taught. Consequently, assessment reform is key to the transformation of the educational system as a whole. It is a "determiner" of learning in two senses. Assessment is the means by which society determines what students have learned and what they can do next. These student assessments are often "high stakes"; test scores certify student achievement, permit advancement or graduation, and determine competitive advantage in further study. High stakes assessments include the SAT, ACT, and

Advanced Placement exams in the U.S., the O-Level (or GCSE) and A-Level exams in most Commonwealth countries, the *Matura* in much of Eastern Europe, and the *Abitur* in Germany, Austria, and Finland.

National assessments are used to determine the effectiveness of teachers, schools, and entire educational systems. These assessments are often also “high stakes”; student performance on tests scores is connected to rewards and punishments for schools and teachers. International assessments are often high stakes for policymakers interested in how their school systems compare with those of other countries. Students, parents, teachers, administrators, and entire schools systems respond accordingly to these high-stakes assessments and it is in this second sense that they have also come to determine what is learned. Whatever the formal curriculum says, whatever teachers are taught to do in their training, whatever it is that students want to learn, the paramount determiner of what is taught, how it is taught, and what is learned is what is assessed, particularly on high-stakes exams. These summative, high-stakes assessments that determine students’ futures, establish rewards and punishments for schools and teachers, and shape classroom and instructional practices of classrooms are the focus of this call to action.

Unfortunately, these traditional assessments do not measure all the competencies and skills that are needed by the 21st century workplace and society (Pellegrino, et, al., 2004). There is a significant gap for assessments, and for the rest of the education system, between what happens in schools and what happens outside of schools (as summarized in Box 1). While people contemporary business work with others and use subject knowledge and a variety of technological tools and resources to analyze and solve complex, ill-structured problems or to create products for authentic audiences, students taking traditional exams do so without access to other people or resources and are, in the main, required to recall facts or apply simple procedures to pre-structured problems within a single school subject.

Standardized Student Assessments	**Tasks in the Outside World**
Assessments are designed primarily to measure knowledge of school subjects and these are divided by disciplinary boundaries.	Subject knowledge is applied within and across disciplinary boundaries along with other skills to solve real world problems, create cultural artifacts, and generate new knowledge.
Students are assessed on their ability to recall facts and apply simple procedures in response to well-defined, pre-structured problems.	People respond to complex, ill-structured problems in the real world contexts.
Students take the exam individually.	People work individually and in groups of others with complementary skills to accomplish a shared goal.
Students take a “closed-book” exam, without access to their notes or to other sources of information, and use only paper and pencil during the assessment.	People use a wide range of technological tools and have access to a vast array of information resources and the challenge is to sort through all of it to find relevant information and use it to analyze problems, formulate solutions, and create products.
Students respond to the needs and requirements of the teacher or school system.	People respond to official standards and requirements and to the needs and requirements of an audience, a customer, or a group of users or collaborators.

This gap between school assessments and the world outside of school fails to prepare students for the demands of the 21st century. As Stanford Professor Linda Darling-Hammond (2005) points out, when high-stakes assessments are emphasized in schools, the use of pedagogical methods focused on the teaching of complex reasoning and problem solving decreases. Teachers report that with such assessments, they have little time to teach anything that is not on the test and that they have to change their teaching methods in ways that are not beneficial to students (Pedulla, et al., 2003). For example, when writing is assessed with paper and pencil, teachers are less likely to use computers when they teach writing (Russell & Abrams, 2004). This is despite the pervasive use of word processors for writing in the real world and the fact that research on the use of word processors consistently shows high levels of impact on the quality of student writing (Bangert-Drowns, 1993; Kulik, 2003).

Traditional assessments also fail to measure all the skills that are believed to be enabled and acquired by the regular use of new, technology-based learning environments. A great deal has been learned about how teachers can integrate the use of ICT into everyday classroom practices and how students can use them to work in teams and to apply their deep understanding of school subjects and ICT tools to solve complex real world problems (Bransford, et al, 2001). For example, international case studies of innovative classrooms (Kozma, 2003) have documented the use of ICT in which students work in groups to specify their own research topics, search the web for related information, use data-loggers to collect science data or web forms to enter survey data, use data bases or spreadsheets to analyze the data, use email to communicate with outside experts, and use word processors, graphics software or presentation software to prepare reports. Video and audio equipment and editing software can be used to create video presentations or performances to be posted on the web and shared with larger audiences. Simulations are used to help students understand complex systems. But traditional assessments do not examine these novel classroom approaches.

However, laboratory studies (e.g. see Bransford & Schwartz, 1999; Schwartz, Bransford & Sears, 2005) show that new approaches to assessments reveal the strengths of innovative pedagogical approaches. A key goal for this project is to examine these classroom innovations and find ways to take ICT-based learning environments and assessments out of laboratories and classrooms, scale them up, and derive implications for international and national high-stakes assessments of 21st century skills and for classroom practices that support assessment reform.

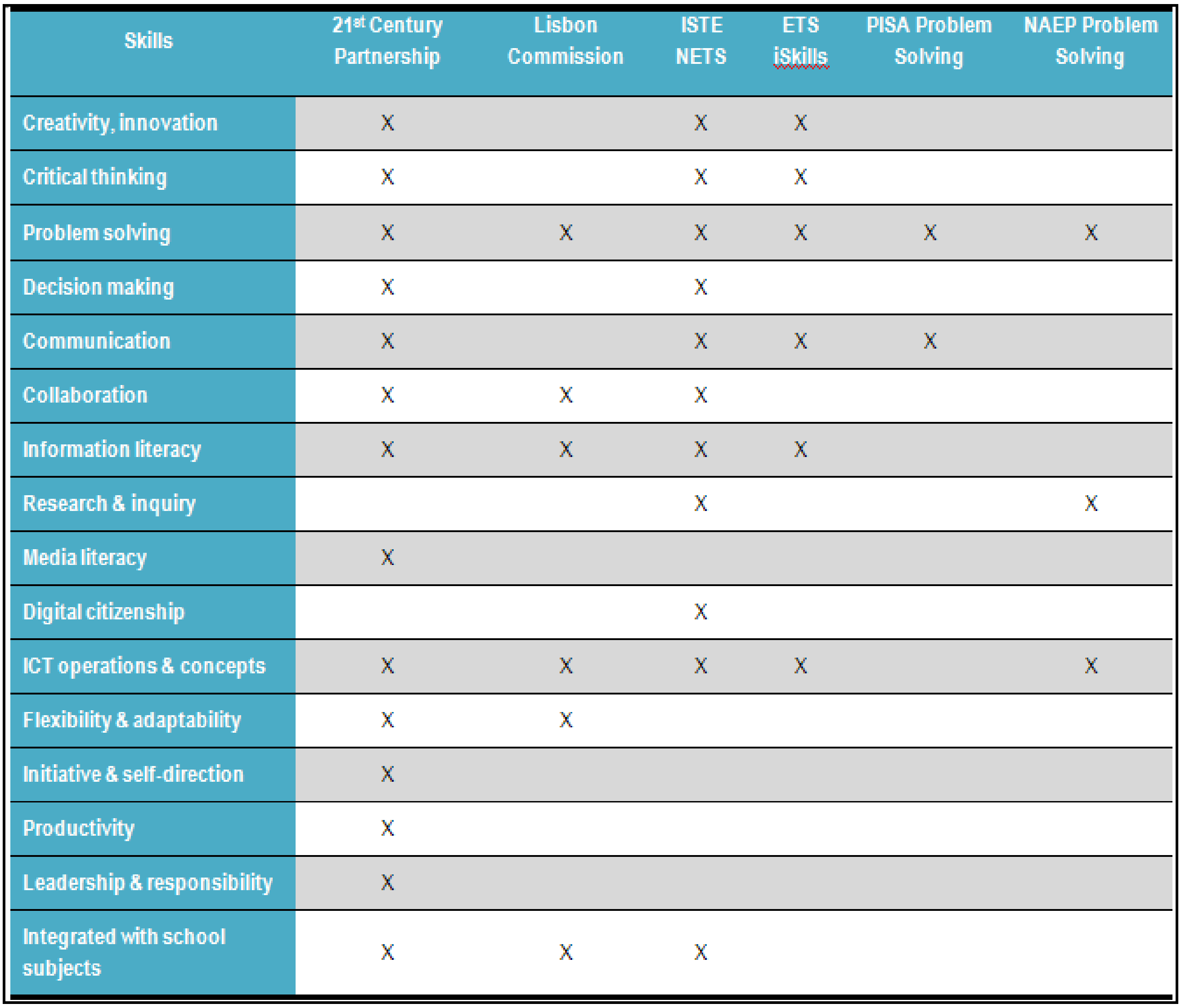
The current state of assessing 21st century skills

Current State of 21st Century Skills Development

A number of high-profile efforts have been launched to identify the skills needs to succeed in the 21st century. Table 1 compares these efforts. Paramount among them is the work of the Partnership for 21st Century Skills (www.21stcenturyskills.org). The Partnership brought together the business community, education leaders, and policy makers to create a vision of 21st century learning and to identify a set of 21st century skills. Built around core subjects, the skills include learning and innovation skills; information, media, and technology skills; and life career skills (See Table 1 for a complete list and comparison). These skills have been adopted by a number of states in the U.S., including Maine, North Carolina, West Virginia, and Wisconsin. Similarly, the Lisbon Council (2007) in the European Union crosses knowledge in science, engineering, mathematics, language, and commerce with “enabling skills” that include: technological skills, informational skills, problem solving, adaptability, and team work. Other efforts have focused in on a more-specialized subset of crucial skills, such as ICT literacy or problem solving.

Some organizations define ICT literacy in very narrow terms as the skills needed to operate hardware and software applications. But others define it more broadly. Prominent among them is the International Society for Technology in Education (ISTE; www.iste.org/), which has defined a set of standards that include technology operations and concepts. They position technology skills in the context of school subjects and a broader set of skills that include creativity and innovation, communication and collaboration, research and information fluency, critical thinking, digital citizenship, and technology operations and concepts. These standards have been adopted by a number of countries and U.S. states. The Educational Testing Service (ETS) iSkills project (www.ets.org/iskills/; Katz, 2007) defines ICT skills as the ability to solve problems and think critically about information by using technology and communication tools and information skills that include defining, accessing, evaluating, managing, integrating, and communicating information and creating new knowledge.

In 2003, a special assessment study of the Programme on International Student Assessment (PISA), a program of the Organization for Economic Cooperation and Development (OECD), defined a skill set related to problem solving skills that included understanding the problem, characterizing the problem, representing the problem, solving the problem, reflecting on the solution, and communicating it to others. ETS designed an assessment of problem solving skills for the U.S. National Assessment of Educational Progress (NAEP) that defined problem solving in terms of the scientific inquiry skills of exploration and synthesis, as well as computer skills.



Skills	21st Century Partnership	Lisbon Commission	ISTE NETS	ETS iSkills	PISA Problem Solving	NAEP Problem Solving
Creativity, innovation	X		X	X		
Critical thinking	X		X	X		
Problem solving	X	X	X	X	X	X
Decision making	X		X			
Communication	X		X	X	X	
Collaboration	X	X	X			
Information literacy	X	X	X	X		
Research & inquiry			X			X
Media literacy	X					
Digital citizenship			X			
ICT operations & concepts	X	X	X	X		X
Flexibility & adaptability	X	X				
Initiative & self-direction	X					
Productivity	X					
Leadership & responsibility	X					
Integrated with school subjects	X	X	X			

Figure 1: Range of skills identified

Table 1 shows the range of skills identified by these efforts. While there are some differences between them, there is significant commonality among them. Based on the examination of this commonality, we propose an initial set of core 21st century skills:

- Creativity and innovation
- Critical thinking
- Problem solving
- Communication
- Collaboration
- Information fluency
- Technological literacy
- Embedded in school subjects

Listing of these skills is relatively easy; operationalizing them is much more difficult. For assessment purposes, skills must be defined precisely and in measurable terms so that assessment tasks can be designed and scoring rubrics can be specified. A key goal of this project is to work with multiple stakeholders to specify these 21st century skills in measurable ways that are relevant to real world work and everyday situations. This will be particularly challenging for skills such as innovation, critical thinking, and collaboration. Specifically, the project will build on previous work in this area to refine the definition of these skills and develop a coherent assessment framework and set of measureable standards for each of the skills.

Current State of Assessment

Many countries have a national assessment of student achievement. Some, such as the Graduate Certificate of Secondary Education (GCSE) and the A-level examinations in the United Kingdom, are taken by all or nearly all students as they progress through their studies. Others, such as the National Assessment of Educational Progress (NAEP) in the United States, test a sample of students for the purpose of measuring the effectiveness of the education system. The major international assessments are PISA, of the OECD, and the Trends in Mathematics and Science Study (TIMSS), of the International Association for the Evaluation of Educational Achievement (IEA). These two assessments differ in that PISA tests 15 year olds and assesses the knowledge in reading, mathematics and science needed to meet the challenges of everyday life of young adults. On the other hand, TIMSS assesses 4th and 8th graders on mathematics and science knowledge that is common to the curricula of participating countries. All of these large scale assessments are focused on the measurement of school subject knowledge, rather than the skills listed above. None currently incorporate the use of ICT tools that are pervasive in the workplace and everyday life.

However some initial efforts have begun to use ICT in the assessment of school subjects. In 2006, 13 countries participated in an optional pilot to test the efficiency and equivalency of delivering science assessment using computers. And in 2009, NAEP will have some computer-based tasks in its science assessment. PISA has the goal of introducing the wider use of ICT in 2009 with the assessment of the reading of electronic texts. PISA is considering the incorporation of ICT in the assessment of mathematics in 2012.

Several projects have begun to explore the use of ICT for the assessment of 21st century skills. In 2003, OECD and ETS conducted a feasibility study that looked at the prospects and difficulties in using ICT to measure ICT literacy skills. ICT literacy was defined as “the ability of individuals to appropriately use digital technology and communication tools to access, manage, integrate, and evaluate information, construct new knowledge, and communicate with others” (Lennon, et al., 2003, p. 8). The two-and-a-half-hour assessment was delivered with ICT and consisted of a multiple choice questionnaire, multiple choice simulated tasks for email, web searching, and database applications, and extended performance tasks involving web search and simulation applications. The assessment was used with a total of 118 students in three participating countries: Australia, Japan, and the U.S. This feasibility study resulted in the development of the ETS iSkills, an assessment of ICT literacy, as well as national ICT Literacy assessment projects in Australia (Ministerial Council on Education, Employment, Training, and Youth Affairs, 2007) and Hong Kong (Law, Yeun, Lee, & Shum, 2007). In Australia, 7,400 students in grades 8 and 10 took an assessment that included the use of both simulated ICT tasks and live applications. In Hong Kong, 2,600 primary and secondary students were assessed on their ICT skills as they used ICT tools in Chinese language, mathematics, and science tasks. In all these assessments, ICT proficiency standards and scoring rubrics were developed and validated. In the U.S., NAEP will assess technology literacy in 2012.

In 2003, ETS conducted a field investigation for the National Assessment of Educational Progress with an ICT-delivered assessment of problem solving in technology-rich environments (Bennett, et al., 2007). The study used two extended scenarios, a search scenario and a simulation scenario, to measure problem solving skills, defined in the context of scientific investigation, and ICT skills. The assessment was given to a nationally representative sample of 2000 8th grade students.

PISA is considering ICT-based assessment of ICT skills in 2012. The IEA is also considering such an assessment of ICT literacy for 2012 or 2014. And in 2012, NAEP is planning to measure technological literacy with an entirely computer-based assessment. A goal of this project is to encourage and support the development of national and international assessments that incorporate the use of ICT.

Beyond problem solving and ICT literacy, SRI’s Center for Technology in Learning developed and pilot tested three ICT-based performance assessments of students’ ability to use various technology tools to access and organize information and relevant data; represent and transform data and information; analyze and interpret information and data; critically evaluate the relevance, credibility and appropriateness of information, data, and conclusions; communicate ideas, findings, and arguments; design products within constraints; and collaborate to solve complex problems and manage information (Quellmalz & Kozma, 2003).

However, due to a variety of methodological and technological barriers, there have been no large-scale implementations of ICT-based assessments of the 21st century skills other than ICT literacy and problem solving. Another goal of this project is to work with multiple stakeholders to promote and support the development of ICT-based assessments for the full range of 21st century skills within the context of school subjects and real world problems. Specifically, this context includes the foundational ideas that organize the factual knowledge of school disciplines and the key questions that make this knowledge relevant to real world situations. The project will use the 21st century skills framework and standards to collect or produce, if necessary, and share examples of ICT-based assessment tasks for each skill and catalog or develop, if necessary, scoring rubrics for skills measured by each task.

Technological and methodological challenges

Technological Advantages, Challenges, and Preconditions

While not all assessment reforms require the use of ICT, technology provides some significant advantages when introduced into assessment. The incorporation of ICT into large-scale assessments promises a number of significant advantages. These include:

- Reduced costs of data entry, collection, aggregation, verification, and analysis.
- The ability to adapt tests to individual students, so that the level of difficulty can be adjusted as the student progresses through the assessment and a more-refined profile of skill can be obtained for each student.
- The ability to efficiently collect and score responses, including the collection and automated or semi-automated scoring of more-sophisticated responses, such as extended, open-ended text responses.
- The ability to collect data on students' intermediate products, strategies and indicators of thought processes during an assessment task, in addition to the student's final answer.
- The ability to take advantage of ICT tools that are now integral to the practice and understanding of subject domains, such as the use of idea organizers for writing, data analysis tools in social science, and visualization and modeling tools in natural science.
- The ability to provide curriculum developers, researchers, teachers, and even students with detailed information that can be used to improve future learning.

The use of ICT in assessments looks something like this:

Students are given a problem scenario in which they are rangers for a national park experiencing a dramatic increase in the population of hares that threatens the ecology of the park. They are asked to decide whether or not to introduce more lynx into the system and, if so, how many. Students receive, respond to, and initiate simulated communications with other rangers who are working on the project and have specialized knowledge of the situation. They search the World Wide Web to find out pertinent information on both hares and lynxes. They organize and analyze this information and evaluate its quality. They make predictions based on their analyses, test their predictions with modeling software, and analyze the results, as represented in graphs, tables, and charts. They integrate these findings with information from other sources and create a multimedia presentation in which they make and defend their recommendations and communicate these to others. (Example courtesy of Edys Quellmalz.)

Such assessments correspond to the situations in the outside world. In the implementation of these assessments, there may be certain local technological barriers that must be overcome related to operating system, hard ware, software, and networking and bandwidth. A goal of this project would be to specify the range of preconditions that might be required of schools to use ICT-enabled learning environments and participate in ICT-based assessments. Among the technological challenges that might inhibit the use of ICT-based assessments are:

- Significant start-up costs for assessment systems that have previously implemented only paper and pencil assessments. These costs would include hardware, software, and network purchases; software development related to localization; and technical support and maintenance.
- The need to chose between the use of "native" applications that would not allow for standardization but would allow students use the applications with which they are most familiar, the use of standardized off-the-shelf applications that would provide standardization but may disadvantage some students that regularly use a different application, or the use of specially developed "generic" applications that provide standardization but disadvantage everyone equally.

- The need to integrate applications and systems so that standardized information can be collected and aggregated.
- The need to choose between stand-alone implementation versus internet-based implementation. If stand-alone, the costs of assuring standardization and reliable operation, as well as the costs of aggregating data. If internet-based, the need to choose between running applications locally or having everything browser-based.
- If the assessment is internet-based, issues of scale need to be addressed, such as the potentially disabling congestion for both local networks and back-end servers as large numbers of students take the assessment simultaneously.
- Issues of security are also significant with internet-based assessments.
- The need to handle a wide variety of languages, orthographies, and symbol systems for both the delivery of the task material and for collection and scoring of open-ended responses.
- The need to keep up with rapidly changing technologies and maintaining comparability of results, over time.
- The need for tools to make the design of assessment tasks easy and efficient.
- The lack of knowledge of technological innovators about assessment, and the corresponding paucity of examples of educational software that incorporates with high-quality assessments.

Methodological Challenges

Significant methodological challenges include:

- The need to determine the extent to which ICT-based items that measure subject knowledge should be equivalent to legacy paper and pencil-based results.
- The need to detail the wider range of skills that can only be assessed with ICT.
- The need to determine the age-level appropriateness of various 21st century skills.
- The need to design complex, compound tasks in a way such that failure on one task component does not cascade through the remaining components of the task or result in student termination.
- The need to integrate foundational ideas of subject knowledge along with 21st century skills in the assessments. At the same time, there is a need to determine the extent to which subject knowledge should be distinguished from 21st century skills in assessment results.
- The need to incorporate qualities of high-level professional judgments about student performances into ICT assessments, as well as support the efficiency and reliability of these judgments.
- The need to develop new theories and models of scoring the students' processes and strategies during assessments, as well as outcomes.
- The need to establish the predictive ability of these judgments on the quality of subsequent performance in advanced study and work.
- The need to distinguish individual contributions and skills on tasks that are done collaboratively.

A key goal of this project is to identify, elaborate on, and address the barriers to ICT-based assessment of 21st century skills and work with partners to develop and implement breakthrough methodologies and technologies.

An Action Plan

In response to the urgent and crucial need for assessment reform to advance educational transformation, Intel, Microsoft, and Cisco have set up a structure and a series of actions to address this need. We are currently identifying a team of international experts that will lead this effort and, with this call to action, invite other interested partners from government ministries, assessment organizations, universities and educational research institutions, foundations, and businesses to join in achieving the challenging goals of this Project.

There are many international and national assessment programs, assessment organizations, NGOs, businesses, research centers, and individual researchers working on the specification of 21st century skills and development of ICT-based formative and summative assessments. The Project will leverage these existing efforts and add value to them for the purpose of transforming educational assessment for the 21st century. Specifically, the Project will add value by catalyzing this international community to identify the problems, issues, and barriers that:

- are common to all,
- that are of the highest priority, and
- cannot be addressed by any individual project alone.

Furthermore, the Project will provide a structure by which this international community can draw on and share existing knowledge and create effective solutions to address the problems, issues, and barriers associated with the identified skills and foster wide-scale adoption of assessment reforms.

Five working groups form the core of an international expert-led project

The goals of the Project will be accomplished by the Executive Director, Dr. Barry McGaw of the University of Melbourne, and a Management Team that is organized into five Working Groups:

1. The 21st Century Skills Working Group, led by Ms. Senta Raizen of WestEd. This group will specify high priority 21st century skills in measurable form.
2. The Classroom Learning Environments and Formative Evaluation Working Group, headed by Dr. John Bransford of the University of Washington. This group will review classroom-based, ICT-enabled learning environments that emphasize interactive, formative assessments and provide opportunities for students to reach important criteria at their own rates, and derive implications from these environments for summative assessments and for classroom practices aligned with assessment reform.
3. The Methodological Issues Working Group, led by Dr. Mark Wilson of the University of California, Berkeley. This group will identify methodological problems and specify solutions for development of assessments of 21st century skills.
4. The Technological Issues Working Group, led by Dr. Beno Csapo of the University of Szeged. This group will identify technological problems and specify solutions for scalable ICT-based assessments of 21st century skills.
5. The Country Deployment Working Group. This group will ensure there is coordination and knowledge-sharing by multiple stakeholders, both within and across partner countries, as well as between countries and the other working groups and between participating countries and the partner companies.

The work of the Project will be organized around a series of annual working conferences, online-interactions, and a knowledge sharing web portal. A public, knowledge-sharing portal will collect and share examples of the measurement specifications of various 21st century skills and assessment frameworks, tasks, and scoring rubrics. The portal will also post the finished works of the Project.

How you can get involved

In the context of the Project's goals, structure, and activities, we are looking for:

- Assessment experts, researchers, business leaders, policymakers, and non-governmental organizations—especially those who have been working in this area—to help identify and specify 21st century skills in measurable ways.
- Assessment experts, researchers, educators, software developers, and ministry officials to develop, collect, and share exemplary ICT-based assessment tasks and scoring rubrics.
- Assessment experts, researchers, and software and network engineers—especially those who have been the leaders in experimenting with ICT-based assessment—to share their experience and expertise, identify and address the barriers to ICT-based assessment, and develop breakthrough technologies and analytic methodologies.
- Policymakers and ministry officials who are interested in having their countries help shape and refine the efforts of the Project and participate in the implementation and pilot testing of the new assessments.
- Businesses, foundations, and government agencies to co-fund these important efforts in private-public partnership.

The project began in January 2009 and will run for approximately three years. Those interested in participating in this effort should contact the Project Executive Director, Dr. Barry McGaw at bmcgaw@unimelb.edu.au.

References

Apte, U., Kamakar, U. & Nath, H. (in press). Information services in the US economy: Value, jobs, and management implications. California Management Review.

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

Bangert-Drowns, R. L. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. Review of Educational Research, 63, 69-93.

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: US Department of Education, National Center for Education Statistics.

Black, S. & Lynch, L. (2003). What's driving the new economy: The benefits of workplace innovation. The Economic Journal, 114, 97-116.

Bransford, J. & Schwartz, D. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & D. Pearson (eds), Review of Research in Education (pp. 61-100). Washington, D.C.: AERA.

Darling-Hammond, L. (2005). The consequence of testing for teaching and teacher quality. Yearbook of the National Society for the Study of Education, Volume 104, Issue 2, Page 289-319

Fallows, D. (2008). Search engine use. Pew Internet and America Life Project. http://www.pewinternet.org/reports.asp, accessed on 7-1-08.

Gera, S., & Gu, W. (2004). The effect of organizational innovation and information technology on firm performance. International Performance Monitor, 9.

Kamakar, U. & Apte, U. (2007). Operations management in the information economy: Information products, processes, and chains. Journal of Operations Management, 25, 438-453.

Katz, I. (2007). Testing information literacy in digital environments: ETS's iSkills assessment. Information Technology and Libraries, September, 1-12.

Kozma, R. (Ed.) (2003). Technology, Innovation, and Educational Change: A Global Perspective. Eugene, OR: International Society for Technology in Education.

Kulik, J. (2003). The effects of using instructional technology in elementary and secondary schools: What controlled evaluation studies say. Menlo Park, CA: SRI International.

Lennon, M., Kirsch, I., Von Davier, M., Wagner, M., & Yamamoto, K. (2003). Feasibility study for the PISA ICT literacy assessment. Princeton, NJ: Educational Testing Service.

Lisbon Council (2007). Skills for the future. Brussels: Lisbon Council.

Law, N., Pelgrum, W. & Plomp, T. (2008). Pedagogy and ICT use in schools around the world: Findings from the IEA SITES 2006 study. Hong Kong: Springer.

Law, N., Yeun, A., Lee, Y., & Shum, M. (2007). Information literacy of Hong Kong students. Hong Kong: University of Hong Kong.

Lenhart, A., Madden, M., Macgill, A., & Smith, A. (2007). Teens and social media. Pew Internet and America Life Project. http://www.pewinternet.org/reports.asp, accessed on 7-1-08.

Ministerial Council on Education, Employment, Training, and Youth Affairs (2007). National assessment program: ICT literacy. South Carlton, Australia: Ministerial Council on Education, Employment, Training, and Youth Affairs.

Ofcom (2008). Social networking. London: Ofcom.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2004). Knowing what students know: The science and design of educational assessment. Washington, D.C.: National Academy Press.

Pilat, D. (2004). The economic impact of ICT: A European perspective. Paper presented at a conference on IT Innovation, Tokyo.

Pedulla, J.J., Abrams, L.M., Madaus, G.F., Russell, M.K., Ramos, M.A., & Miao, J. (2003). Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers. Boston: National Board on Testing and PublicPolicy, Boston College.

Quellmalz, E. & Kozma, R. (2003). Designing assessments of learning with technology. Assessment in Education, 10(3), 389-407.

Rainie, L. (2008). Video sharing websites. Pew Internet and America Life Project. http://www.pewinternet.org/reports.asp, accessed on 7-1-08.

Russell, M., & Abrams, L. (2004). Instructional uses of computers for writing: The impact of state testing programs. Teachers College Record, 106(6), 1332–1357.

Schwartz, D.L., J.D. Bransford, and D. Sears, (2005). Innovation and efficiency in learning and transfer. In J. Mestre (ed.), Transfer of Learning from a Modern Multidisciplinary Perspective, pp. 1–51, Mahwah, NJ: Erlbaum.

Zohgi, C., Mohr, R., & Meyer, P. (2007). Workplace organization and innovation (Working Paper 405). Washington, D.C.: Bureau of Labor Statistics.

The author:

Robert B. Kozma, Ph.D.
Principal Consultant
Kozmalone Consulting
Technology in the Service of Development
2151 Filbert St.
San Francisco, CA 94123 USA
http://robertkozma.com

Dr. Robert Kozma is an independent consultant and Emeritus Director and Principal Scientist at the Center for Technology in Learning at SRI International in Menlo Park, California, where he was active for 10 years. For 20 years prior to that, he was at the University of Michigan as a professor and a research scientist. His expertise includes ICT policy that links education reform to economic and social development and he has consulted with Ministries of Education numerous countries, as well as with Intel Corporation, Cisco, Microsoft, the World Bank, OECD, UNESCO, the IEA.

The European Coherent Framework of Indicators and Benchmarks and implications for computer-based assessment
History, issues and current status

Oyvind Bjerkestrand
European Commission

Abstract:
Indicators and benchmarks are key elements of evidence-based policy making and the monitoring of progress of the Lisbon process. There is a rapidly increasing interest in getting internationally comparable data at national and international levels. Recent development in education and training has increased the demand for indicators and especially in order to measure skills and competences. For some educational areas large-scale international surveys already exist, and others are under preparation. The Commission is following closely the surveys developed by the OECD, IAE and other international organisations. The Commission is also launching future surveys to collect data in different fields. In order to minimise the burden for the participating countries and the participants in international surveys the Commission is eager to develop further the possibilities of using computer-based tests and assessments when launching surveys where this can be effective and efficient, and enhance the quality of the tests, assessments and the publication of results.

Lisbon agenda

The Lisbon Strategy was launched by the European Council in Lisbon in 2000 and its aim is to make EU "the most dynamic and competitive knowledge-based economy in the world capable of sustainable economic growth with more and better jobs and greater social cohesion, and respect for the environment by 2010" (European Council 2000). The strategy rests on an economic pillar preparing the ground for the transition to a competitive, dynamic, knowledge-based economy, a social pillar designed to modernise the European social model by investing in human resources and combating social exclusion as well as an environmental pillar.

Drawing on lessons learnt from five years of implementing the Lisbon strategy, the European Council in 2005 relaunched the strategy. It agreed to give priorities to jobs and growth and sought a stronger mobilisation of all appropriate national and community recourses. The revised strategy places strong emphasis on knowledge, innovation and the optimisation of human capital. Education and training are critical factors to develop EU's long term potential for competitiveness as well as social cohesion.

In its annual report "*Progress towards the Lisbon objectives in education and training"* (European Commission, 2004-2008) the Commission is examining the performance and progress in education and training systems in the EU under the Education and Training 2010 work programme. The purpose of this series of reports is to draw on indicators and benchmarks in order to provide strategic guidance for the work programme and to set out the evidence available on progress towards the objectives agreed by ministers.

This article is based on the progress reports and concentrates on the development of new indicators in the coherent framework on indicators and benchmarks and the implications for using computers for testing pupils' skills.

Five European Benchmarks for 2010

Regular monitoring of performance and progress using indicators and benchmarks is an essential part of the Lisbon process. The *Open method of coordination* in education and training require tools for measuring progress and identifying good performance and practices in Member States. Indicators and benchmarks serve as tools for evidence-based policymaking at European level. The five benchmarks adopted by the Council in May 2003 are of continuing relevance in guiding policy action and they are central tools in the follow up of the Lisbon process. By adopting five European benchmarks, the Council set measurable objectives indicating the policy areas in which it expected to see clear progress. The benchmarks to be achieved by 2010 are:

- No more than 10% early school leavers;
- Decrease of at least 20% in the percentage of low-achieving pupils in reading literacy;
- At least 85% of young people should have

completed upper secondary education;

- Increase of at least 15% in the number of tertiary graduates in Mathematics, Science and Technology (MST), with a simultaneous decrease in the gender imbalance;
- 12.5% of the adult population should participate in lifelong learning.

From data gathered since 2000 we can observe that education and training systems in the EU are generally improving. Although there is progress, attaining the benchmarks set for 2010 will need more effective initiatives across the Member States. Four of the five benchmarks show progress, and the benchmark on mathematics, science and technology graduates was reached already in 2003. The situation for reading literacy of young people is getting worse; the share of low achievers in reading literacy has increased from 2000 to 2006 (See figure 1).

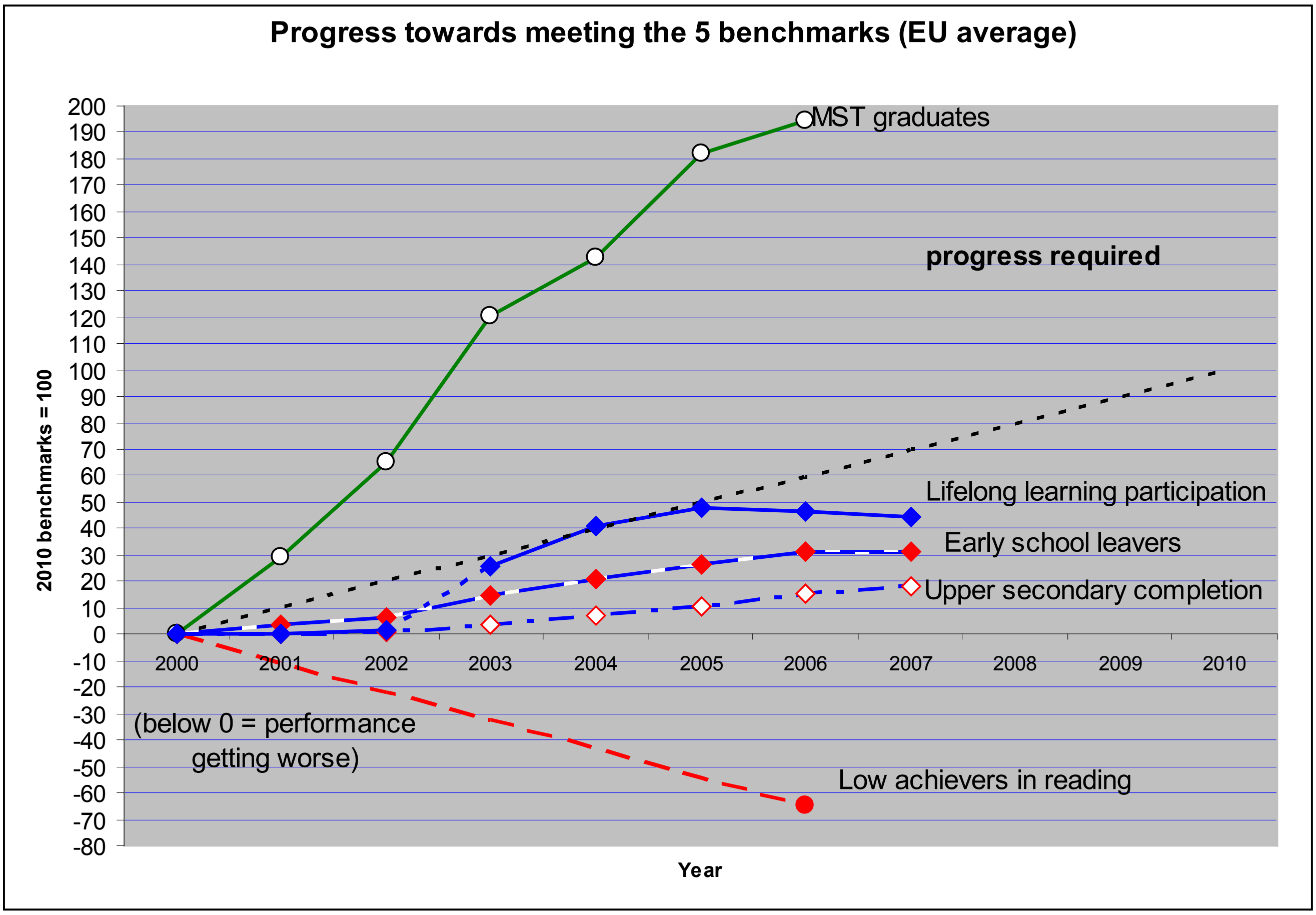


Figure 1: Progress towards the five benchmarks for 2010 (EU average)

Key competences

Acquiring basic skills is essential for functioning in the rapidly changing and highly interconnected knowledge based society. Acknowledging the importance of acquiring basic skills the European council in 2002 underlined the need to improve the mastery of basic skills by adopting a resolution on Lifelong learning and "the new basic skills".

A *Recommendation of the European Parliament and the Council on key competences for Lifelong learning* was published in December 2006 (Council 2006a).

In this recommendation it was stressed that “*As globalisation continues to confront the European Union with new challenges, each citizen will need a wide range of key competences to adapt flexibility to a rapidly changing and highly interconnected world*".

The *Recommendation of the European Parliament and the Council* defined a reference framework with a combination of knowledge, skills and attitudes which all individuals need for personal fulfilment and development, active citizenship, social inclusion and employment.

The reference framework consists of eight competences:

1. Communication in the mother tongue;
2. Communication in foreign languages;
3. Mathematical, science and technology competence;
4. Digital competence;
5. Learning to learn;
6. Social and civic competences;
7. Sense of initiative and entrepreneurship;
8. Cultural awareness and expression.

The eight competences are considered as equally important. Several competences that are defined in the framework, like social and civic competences, sense of initiative and entrepreneurship, learning to learn, and cultural awareness and expression are not only learned in the traditional education at school, but require new approaches in organising learning. Teachers need to work together with each other, with the local community and deal with heterogeneous groups. This also poses challenges for testing these skills.

Data are already available for some of the key competences, while for others surveys will have to be launched in order to collect data. Future rounds of existing surveys, like the PISA survey, will yield updated data for indicators on pupils' skills in reading, mathematics and science.

Coherent Framework of Indicators and Benchmarks

In May 2007 the Education Council adopted conclusions on a coherent framework of 16 core indicators for monitoring progress towards the Lisbon objectives in education and training (Council 2007). The indicators are listed in table 1.

1)	Participation in pre-school education
2)	Special needs education
3)	Early school leavers
4)	Literacy in reading, mathematics and science
5)	Language skills
6)	ICT skills
7)	Civic skills
8)	Learning to learn skills
9)	Upper secondary completion rates of young people
10)	Higher education graduates
11)	Cross-national mobility of students in higher education
12)	Participation of adults in lifelong learning
13)	Adults' skills
14)	Educational attainment of the population
15)	Investment in education and training
16)	Professional development of teachers and trainers

Table 1: Core indicators

The coherent framework permits the Commission to underpin key policy messages, to analyse progress, to identify good practise and to compare the performance with third countries.

For some of the core indicators in the coherent framework data already exist. For other indicators data are in the phase of being developed or new surveys are needed in order to get the data. The framework is mostly covered by statistical data that already exist and which have been used in monitoring the follow-up of the Lisbon objectives in education and training. These indicators are continuously being improved within their specific statistical infrastructures: European statistical system (ESS), UNESCO/OECD/EUROSTAT (UOE) data collection and OECD/PISA survey. However, in the case of the five core indicator areas, mainly concerning the key competences, new data needs to be collected.

It is evident that some of the areas covered in the *Framework of Key Competences* are covered by the coherent framework. These are: literacy in reading, mathematics and science, language skills, ICT skills, civic skills and learning to learn skills.

Development of new indicators/surveys

Five cross-national surveys will be implemented in the next couple of years in the core indicators' areas demanded by the Council. The planned schedules for the presentation of the results from these surveys are from 2008 to 2013. For two of the core indicators, new surveys are presently being prepared by the European Union: For the core indicator on "Language skills" a European survey is being implemented and final results are expected in 2012. For the "Learning to learn skills" a pilot survey is presently ongoing and results are expected in 2008. In the case of the three other core indicator areas, new surveys are implemented in cooperation with other international organisations. In the areas of "Adult skills" and "Teachers professional development", EU data needs can be satisfied within new surveys organised by the OECD. The presentation of the results from the Programme for the International Assessment of Adult Competences survey (PIAAC) is planned for 2013 and the Teaching and learning international survey (TALIS) is presently being implemented, and results are foreseen in 2009. For the core indicator on "Civic skills" a European module has been included in the on-going International Civics and Citizenship Education Study (ICCS) prepared by the International Association for the Evaluation of Educational Achievements (IEA). The survey is presently being implemented and results are foreseen in 2009.

It is important for the Commission that all countries following the Lisbon process will participate in the new surveys. A European indicator based on data from few countries would be of lesser quality and would not be able to play its full role as a tool for monitoring progress and identify good performances. The new surveys will not only give valid and comparable data for the development of core indicators but also provide extensive contextual data and information which will make it possible to carry out secondary analysis producing new knowledge about learning processes in these fields.

The use of computer-based assessment in testing Language skills

Improving language skills in Europe is an important objective as part of the Lisbon growth and jobs strategy. The Barcelona European Council expressed interest in language learning when it called for "the mastery of basic skills, in particular by teaching at least two foreign languages from a very early age." As mentioned earlier, knowledge of foreign languages is now recognised as one of the key competences and part of the coherent framework.

Since there is no existing standardised survey of language skills across the Union, it is necessary to collect accurate and up-to-date data on the effectiveness of foreign language teaching systems. The European Language Indicator will illustrate the general level of foreign language knowledge of the pupils in the Member States.

The Council conclusions on the European Indicator of Language Competence asked for measures for objective testing of skills in first and second foreign languages (Council 2006b). The Council invited the Commission to assist the Member States to define the organisational and resource implications for them and to take *"into consideration the need to prevent undue administrative and financial burdens for the Member States"* (Council 2006b). Furthermore, the Council invited the Commission to set up an Advisory Board of national experts to advise the Commission on the preparation and implementation of the tests. The Commission, together with the Advisory Board, were asked to take into consideration the preferred administrating of the tests, and the possibilities of e-testing. The idea is that in general terms electronic testing could improve the effectiveness, i.e. improve identification of skills, and efficiency, by reducing costs (financial efforts, human resources etc.).

International surveys are expensive to develop. On one hand it is the international costs in preparing the framework, the test items, and the organisation of the survey. On the other hand the most expensive part is the implementation of the tests in the participating countries. The level of national costs depends on several factors, for example the testing methods used, the sample size, the number of skills tested, the setting up and running of organisational support structures, training of national coordinators, quality assurance procedures etc. Hence the increased demand for both national and international large-scale assessments makes the possible benefits from computer-based tests and assessments highly relevant. Several countries have already introduced computer-based national tests, and these experiences could be used in order to introduce computer-based assessments at the international level.

There are no doubt both challenges and benefits when introducing computer-based assessment and computer-based testing. The discussion in the Advisory Board identified both positive and negative aspects by introducing computers for testing. The positive points that have been identified are for instance the fact that is an efficient way of testing; the marking and coding as well as statistical treatment of results can be much faster and error proof; the collection of data and the use of adaptive testing could be easier, new test concepts can be introduced and the data processing costs could be reduced. However, the Board identified a variety of aspects to be taken into account when using computers for testing, such as software quality and compatibility, the risk of testing candidates that are not used to the tool, hence testing other skills than intended; typing problems; and possible high investment costs. Many experts agree on the overall added value and advantages of e-testing in large-scale assessments.

In general terms the introduction of computer-based testing would be the optimal step forward in relation to the language survey. However, there are different levels of readiness in the participating countries concerning testing with computers. The Commission published a call for tender where the tenderer should provide a survey based on alternative or complementary testing based on computers and on paper and pencil tests. The realisation of the survey was attributed to the consortium SurveyLang and they are developing the testing instrument to be realised for the survey in 2011.

The Commission is supporting the work being done in this field at the Centre for Research on Lifelong learning (CRELL). CRELL has taken a number of initiatives in this respect, for example:

- To carry out comparative research on modes and platforms for computer-based testing and to analyse effectiveness of software implementations for large-scale surveys;
- Explore the state-of-the art in European education and focus on large-scale assessments case studies;
- The project includes a pilot survey in cooperation with the Ministry of Education in Iceland on on-going national school curriculum assessment activities in: Reading (using sequential computer-based and computer adaptive testing), Mathematics, English (computer-based and adaptive).

In this sense; it remains a huge challenge to run international tests. The developmental work in this field is extremely important for the European Commission in view of launching new international surveys. In the future development of new surveys the use of computers will be considered also in other fields than the European Survey on Language Competences. The experience learned from this survey, together with the work being done in OECD and IEA will help the progress in this field that ultimately all participants and developers of tests can benefit from.

References

European Council (2000). Precidency Conclusions. Lisbon European Council 23 and 24 March 2000.

European Commission (2004-2008) Progress Towards the Lisbon Objectives in Education and Training. Indicators and Benchmarks. Commission Staff Working Paper. Annual report 2004-2008.

Council (2007), A coherent framework of indicators and benchmarks for monitoring progress towards the Lisbon objectives in education and training. Council conclusions of 25 May 2007, 2007/C1083/07

Council (2006a), Key Competences for Lifelong learning, Recommendation of the European Parliament and of the Council of 18th December 2006, 2006/962/EC

Council (2006b). Conclusions on the European Indicator of Language Competence. Council conclusion of 25 July 2006, (2006/C 172/01)

The author:

Oyvind Bjerkestrand
European Commission –
Directorate General for Education and Culture
Directorate A – Lifelong Learning: horizontal Lisbon policy issues and international affairs
Unit 4: Analysis and studies
MADO 8/14
B-1049 Brussels
Belgium

E-Mail: oyvind.bjerkestrand@ec.europa.eu

National expert in the European Commission, Directorate-General for Education and Culture, Directorate "Lifelong learning: horizontal Lisbon policy issues and international affairs, Unit A4 "Analysis and Studies". His work concentrates on supporting and developing the implementation of lifelong learning policy, with particular regard to the key role of education and training in the Lisbon strategy. Bjerkestrand is working in the field of developing new indicators for monitoring progress in the field of education and training towards the Lisbon objective.

Computer-based Assessment and the Measurement of Creativity in Education

Ernesto Villalba
Joint Research Centre, IPSC, CRELL (Italy)

Abstract:
The European Council agreed in declaring 2009 the year of creativity and innovation. The Communication of March 2008 (European Commission, 2008, 159 final) puts it simply: "Europe needs to boost its capacity for creativity and innovation for both social and economic reasons". The promotion of creativity brings with it the necessity of assessing it, at least partially, in order to determine if policies are effective. Measures of creativity have been used in psychological studies with more or less intensity in the last 50 years or so. Creativity is also at the core of an emerging body of literature in other areas such as management (Nonaka, 1991) and urban policies (Florida, 2002).Creativity is a vague entity, difficult to define and measure. However, certain consensus exists in some of its characteristics. It seems clear that it is related to something new and with some short of value. There is also consensus on the idea that everybody can be creative to some extent. With the raise of Computer-based Assessment that permits more sophisticated test items, the question is: Will it be possible to measure creativity in an international comparative way? The paper presents a short review of the different ways used to measure creativity and concludes with some reflections of the possibilities of measuring creativity using existing large scale assessment tools, specifically computer-based assessment.

Introduction

This paper presents the point of departure of a challenge posted to experts on computer-based assessment, specifically: Will it be possible to measure creativity? The paper aims at constructing a preliminary understanding of research on creativity. It explores the possibilities of constructing a "creativity indicator" using large scale surveys and how computer-based assessment will allow for measurement of creativity that are not currently available. Thus, the paper is mainly focused on exploring measurement possibilities of creativity in an international comparative manner.

The paper aims at answering (at least partially) the following questions:

- Is it possible to measure creativity in a comparative international manner?
- Is it possible to use existing large scale surveys to assess creativity?
- Is computer-based assessment the answer to the problem of measuring creativity in an international comparative way?

The concept of creativity has gained importance in the last years. A vast amount of management literature has been increasingly focusing on how to enhance creativity in the workplace, in order to cope with constant changing environments (see e.g. Nonaka and Takeuchi 1995, Villaba 2008). One sign of the importance of creativity is the decision of the European Union of making 2009, European Year of Creativity and Innovation. Within this frame the project "education for innovation and innovation for education" at the Centre for Research on Lifelong Learning (CRELL) of the Joint Research Centre of the European Commission aims at finding possible ways of measuring creativity in an international comparative manner.

It is clear that the phenomenon of creativity is extremely complex. The study of creativity has different perspectives and approaches. This paper review some work on creativity in psychology, mainly through the extensive reviews of Sternberg and Lubart (1999) and Runco (2007). Section 2 presents an overview of some definition and main issues in relation to creativity. Section 3 reviews different ways that have been used to approach the measurement of creativity in psychology. Finally, section 4 discusses possible models to measure creativity and how computer-based assessment could provide a key factor to advance complex construct assessment.

Towards an understanding of creativity

Wehner, Csikszentmihalyi and Magyari-Beck (1991, 270) have illustrated the problem of the research on creativity with the fable of the blind men trying to describe an elephant by touching different parts of the animal, where the one touching the tail says it is like a snail and other touching the flank says it is like a wall. Sternberg (2006a) found five commonalities in the research of creativity. First, creativity "involves thinking that aims at producing ideas or products that are relatively novel and that are, in some respect, compelling" (Sternberg, 2006a, 2). Second, creativity has some domain-specific and domain-general elements. That is to say, it needs some specific knowledge, but there are certain elements of creativity that cut across different domains. Third, creativity is measureable, at least to some extent. Fourth, it can be developed and promoted. And fifth, "creativity is not highly rewarded in practice, as it is supposed to be in theory" (Ibid.).

Sternberg and Lubart (1999) propose that the origin of creativity research is on spirituality. In this way, research associated with creativity has not had the necessary scientific back-up. Later, "pragmatic approaches on creativity" have been mainly concerned with the development of techniques to promote creative thinking in organizations. They are mainly practical approaches and do not provide a clear idea of what are the characteristics of creativity. Studies on cognitive psychology have tried to understand the process of creative thinking. The debate is centered on delineating creative thinking (Plsek, 1997). Many studies in cognitive psychology assume, however, that creativity is just extraordinary results of ordinary processes (Smith, Ward and Finke, 1995). In other cases some authors maintain that creativity is not much different from intelligence (Getzels and Jackson 1962). Later research seems to agree that intelligence and creativity are related, but only at a certain level of ability (Runco and Albert, 1986, Runco, 2007, 7). Finally, psychometric approaches to creativity have been mainly focused in developing tests to measure creativity. Plucker and Renzulli (1999) differentiate four areas where psychometric methods have been applied in creativity research: (1) creative process, (2) personality and behavioral correlates, (3) characteristics of creative products, and (4) attributes of creative fostering environments. The psychometric approach will be treated in more detailed later on.

Defining creativity

Sternberg and Lubart (1999, 3) maintain that "Creativity is the ability to produce work that is both novel (i.e. original, unexpected) and appropriate (i.e. useful concerning tasks constrains)". Several definitions of creativity maintain that creativity involves the production of something new and useful (Bailin 1988, Bean 1992, Solomon, Powell and Gardner 1999, Mumford 2003, Andreasen 2005 and Flaherty 2005). Runco (2007, 385) maintains that these are product bias definitions. For him, product bias consists on assuming that all creativity requires a tangible product: "It would be more parsimonious to view creative products as inventions and the process leading up to them as creative or innovative" (ibid.).

In the UK, the National Advisory Committee on Creative and Cultural Education (NACCCE) published in 1999 a report where they provided a more elaborated, but similar definition of creativity. They maintain that creativity processes have four characteristics:

(1) Creativity always involves imagination: it is the process of generating something original.
(2) Creativity is purposeful: it is imagination put into action towards an end.
(3) It produces something original in relation to one's own previous work, to their peer group or to anyone's previous output in a particular field.
(4) Finally, creativity has value in respect of the objective it was applied for. Creativity involves not only the generation of ideas, but also the evaluation of them, and deciding which one is the most adequate one.

The NACCCE maintain that they understand creativity in a "democratic" way. That is to say, all of us are somehow creative (NACCCE, 1999, 30). In the UK, the Department for Culture, Media and Sport (DCMS) uses this definition in their plan for actions to enhance creativity in schools in 2006 (DCMS, 2006).

NACCCE (1999) opposed their view on creativity to two other different views: An "elite", and a "sector" definition. An "elite" definition involves that creative people are those with "unusual talents". This relates to the differentiation in the creativity research usually between eminent-level and non-eminent-level creativity. Richards (1999a) defines eminent-level creativity as the one that involves discoveries that are of particular importance for society (for example, scientific discoveries), while, the later refers to the capacity of people to adapt to new situations. The later is in line with the "democratic" understanding of creativity. The "sector" definition maintains that creativity is something associated with the arts, and that it does not involve other sectors of production such as science or technology. An extended version of the sector definition could be found in the definition of creativity that KEA European Affairs (2006) propose. They define creativity in a cross-sector and multidisciplinary way, mixing elements of ‘artistic creativity’, ‘economic innovation’ as well as ‘technological innovation’. Creativity is defined as “a process of interactions and spill-over effects between different innovative processes” (KEA European Affair, 2006, 41). They differentiate between: Scientific, technological, economic and cultural creativity.

Also from a more macro-level approach, Richard Florida’s popular book “The rise of the creative class” provides a view of what creativity encompasses at a societal level (Florida, 2002). Florida’s main thesis is that creativity is the “ultimate economic resource” (Florida, 2004, xiii). He maintains that we live in a “Creative Age”. He is specifically interested in the factors associated with urban economic growth. Florida maintains that creative people are attracted to places that are characterized by a “culture that’s open-minded and diverse” (Florida, 2004, xvii). And it is this creative class the one that has strong influence in making a region prosper economically. In his view, “places provide ecosystems that harness human creativity and turn into economic value” (Florida, 2004, xix). Inspecting the characteristics of these places he presents his 3 T’s model, centered on three main areas: Technology, Talent and Tolerance. For him, these three T’s constitute the main magnets for creative people to establish themselves in a city. He does not provide a specific definition for creativity, but from his description of what are the “main themes” in the body of literature about creativity, one can find a short of definition of creativity (Florida, 2002, 30). In his view, creativity is an essential feature of our life. He presents a “democratic” conception of creativity in line with that of NACCCE, where creativity is embodied in different areas of human life. For him, creativity is multidimensional and experiential. Creativity requires “work” to appear and is usually guided by intrinsic rewards.

Unresolved issues on defining creativity

There is, thus, certain consensus on some of the creativity characteristics. It seems clear that it is related to something new and with some short of value. It also seems that there is certain agreement that everybody can be creative to some extent. However, as Mayer (1999, 450) addresses there are several clarifying questions for which authors on creativity have different answers.

As Mayer (1999) noted, studies on creativity can refer to *personal* or *social* creativity. Personal creativity refers to creating something new in respect to the person that creates the product. Creativity that is social refers to something new and useful in respect to the social or cultural environment where it is produced. NACCCE (1999) maintain that creativity involves originality in three possible ways: *Individual*, *relative* or *historic*. *Individual* creativity coincides with Mayer’s definition of personal creativity. *Relative* refers to originality in relation to their peer group. Finally, *historic*, refers to original in terms of anyone’s previous output in a particular field. Sternberg (1999) proposes that creative contributions in science can be grouped in three major categories: Contributions that (1) accept current paradigms, (2) reject current paradigms and (3) attempt to integrate multiple current paradigms.

In addition, Mayer (1999) maintains that there is a need to clarify if creativity is a property of: (1) *People*, (2) *Product* or (3) *Processes*. He maintains that depending on this assumption, different approaches have been used to study creativity. Runco (2007) adds approaches to creativity related to *place* (Rodhes, 1962; Richards, 1999b; Runco 2004), *persuasion*, in studying how creative people change the way other people think (Simonton, 1990), and *potential* (Runco, 2003) emphasizing research on those that have potential for creativity but are not realizing it.

Sternberg and Lubart (1999) overcome some of these difficulties by advocating for “confluence approaches”. This line of research put together multiple views on creativity, where different components must converge for creativity to occur (see e.g. Amabile 983, Gruber and Davis 1988, Csikszentmihalyi 1996). Their own “investment theory of creativity” (Sternberg and Lubart, 1991, 1992, 1995, 1996) is an example of these confluence approaches. The basic idea is that “creative people are the ones who are willing and able to ‘buy low and sell high’ in the realm of ideas” (Sternberg, 2006b, 87). According to this theory creativity requires six distinct but interrelated resources: intellectual abilities, knowledge, styles of thinking, personality, motivation and environment. Sternberg and Lubart (1999) describe a complex system, where these different resources have to have a proper balance. An example of the complexity can be seen in the case of *knowledge.* Sternberg (2006b, 89) maintains:

"On the one hand, one needs to know enough about a field to move it forward [...] On the other hand, knowledge about a field can result in a closed and entrenched perspective". The rest of the six resources also require the right balance of attributes.

If our aim is to measure creativity in an international, comparative manner, these issues have to be addressed in a measurement model. A working definition of creativity delineating what it is and what it is not is the first and crucial step. An understanding of creativity will necessarily require reflection of the issues presented above. A "confluence approach" will obviously present tremendous challenges, especially in the case of measurement. Measuring complex constructs will not be enough, but it will be necessary to determine what combination of level in these constructs results in creativity. Computer-based assessment will allow for a higher degree of flexibility in measuring some of these aspects, and finding the adequate levels.

Since the interest of the paper is to assess the possibilities of measuring individual-level creativity, the following section reviews main approaches of creativity measurement, mainly in psychology. The review does not claim to be exhaustive, nor comprehensive, but it aims at providing a general overview in the field.

Different approaches to measuring creativity

This section reviews measurements of creativity in the field of Psychology. In most of the cases the approach consists on developing tests to measure creativity. Haensly and Torrance (1990) identified more than 200 instruments for measuring different aspects of creativity. Houtz and Krug (1995) provide a review of several test developed for the assessment of creativity. They followed Hocevar (1981) classifications into: tests of divergent thinking, attitude and interest inventories, personality inventories, biographical measures, ratings by teachers, peers or supervisors, product judgments, self-reports of creative achievements, and eminence or the study of well-known and establish creative people.

Divergent thinking

Houtz and Krug (1995) present the Torrance Test of Creative Thinking, the Wallach and Kogan Tests (Wallach and Kogan, 1965), and the Guilford Battery (Guilford, 1962, 1971) within the category of divergent thinking. Divergent thinking requires open-ended questions; as opposed to convergent thinking problems that always has one or very few correct or conventional answers. McCrae (1987) defines divergent thinking as the ability to generate many difference possibilities for solving a problem. A typical item to test divergent thinking would be to ask to say as many uses of a brick as possible. It is somehow, base on the ideas from associative theories (Mednick, 1962) that maintain that original ideas tend to be remote; they come later in the process of thinking about associations. Creative thinking differs from divergent thinking in that it involves also sensitivity to problems and requires redefinition abilities (Kim, 2006, 4).

The Torrance Test of Creative Thinking (TTCT) is the most widely used test on creativity and has the most extended research on their reliability and validity (Houtz and Krug, 1995, Kim, 2006). The TTCT was developed in 1966, and it has been re-normed four times: 1974, 1984, 1990 and 1998 (see Kim 2006 for a review of the TTCT). Each test pertains to measure: Fluency (The number of ideas), originality (the rarity of ideas), elaboration (the number of added ideas), and flexibility (number of categories of the relevant responses). In 1990 Torrance deleted the flexibility scale, since it correlated highly with fluency (Herbert et al. 2002), and added two norm-reference measures of creative potential: Abstractness of titles and resistance to premature closure (Ball and Torrance, 1980). Abstractness of titles refers to the "degree beyond labelling. It measures the degree a title moves beyond concrete labelling of pictures drawn" (Kim, 2006, 5). Resistance to premature closure pertains to measure the degree of psychological openness. The test can be administered in around 30 minutes, but the process of scoring requires some training and specific country norms. In 1998 the manual provides norms for the United States and includes both grade-related and age-related norms (Kim, 2006). Thus, there is some country specificity in the measurement of creativity. Kim (2006) reported some normative measures in other countries. These norms have usually been developed for research activities.

Heausler and Thompson (1988) refer to four main criticisms regarding the TTCT: Firstly, different order in the presentation of the items leads to different results (Lissitz and Willhof, 1985). Secondly, “creativity tests administered under different conditions lead to differences in performance” (Hattie, 1977, 97). Thirdly, Rosenthal et al. (1983, 39) found that “two raters may agree that a particular student’s performance is better than that of all other students, yet still assigned significantly different scores to describe this performance”. This means as Heausler and Thompson (1988, 464) have pointed out that “these differences might be of practical importance in studies testing mean differences across experimental or other groups”. Finally, a fourth group of criticism refers to the structure of the test. Some studies with factor analysis have shown that the factors found in the TTCT described a task more than underlying constructs (Plass, Michael and Michael, 1974, 413).

Creative personality

Another line of measuring creativity is related to study individual differences and personality attributes. Studies in this line have tried to find characteristics of creative people. They could be divided into psychometric, biographical and historiometric approaches.

In psychometrics approaches, studies attempt “to measure facets of creativity associated with creative people” (Plucker and Renzulli, 1999, 42). Tools in this area for studying creativity consist of lists of personality traits, self-report adjectives check-list, biographical surveys and interest and attitudes measures. The most widely used check list is the Gough’s (1952) Adjective Check Lists (ACL). It consists of 300 descriptor words that a person checks as being self-descriptive. Using such tool, a sample of people that were evaluated as creative by experts is usually compared to other non-creative sample of people. Domino (1970) identified 59 of those descriptors that formed a Creativity Scale (Houtz and Krug, 1995). Other similar tests have been developed and tested with different professionals. Kathena and Torrance (1976) developed the Creative Perception Inventory, composed of the Something About Myself (SAM) and What Kind of Person Are You (WKOPAY) scales. SAM asks people to answer if they have engaged in specific activities with creative potential. It also asks individuals to “agree of disagree with certain self-descriptors, such as ‘I am talented in many different ways’” (Houtz and Krug, 1995, 279). The WKOPAY ask people to check personality traits that they think characterized them.

Biographical and historiometric approaches are mainly related to the study of creative individuals and their context. Biographical approaches involve case studies of eminent creators “using qualitative research methodologies” (Plucker and Renzulli, 1999, 38, see also Gruber and Wallace 1999, for a review). Historiometric is also mainly concerned with the study of eminent creators, names that have “gone down in history” as Simonton (1999) puts it. Through a quantitative analysis of the biographical and historical records related to these eminent creators, historiometrics attempt to measure creativity. The definition of historiometric can be broken down in three components (Simonton, 1999, 117): (1) In historiometric approaches “the goal is the discovery of general laws or statistical relationships that transcend the particular of the historic records” (ibid.). (2) It uses quantitative analyses. The researcher has to transform the usually rich, ambiguous, and qualitative facts of history into more precise, clear numerical measurements. In addition, the researcher should use statistical techniques, mainly multiple regression, factor analysis and latent-variable models to understand the relationship between different aspects of creativity. (3) The subject of study in the historiometric approach is always a “historical individual”.

Runco (2007), in his review of the literature, maintains that creative personality encompasses: Autonomy, flexibility, preference for complexity, openness to experience, sensitivity, playfulness, tolerance of ambiguity, risk taking or risk tolerance, intrinsic motivation, psychological androgyny, self-efficacy and wide interest and curiosity. He also noted that creative personality varies from domain to domain, and perhaps, even from person to person: “there is no one creative personality” (Runco, 2007, 315). He however, maintains that certain characteristics depend on values, intentions and choice; thus, people have the possibility of trying to enhance their creativity or not. This is in line with what Sternberg affirms (2006b, 93) in his review of creativity research: “Creativity is as much a decision about and an attitude toward life as it is a matter of ability”.

Measuring creativity using large international surveys

The previous section has shown that despite having some measures of creativity, it seems clear that understandings of creativity differ depending on the approach chosen to study it. Tests of creativity at individual level require, mainly, either some type of divergent thinking (as opposed to convergent thinking) or some personality traits (that have been associated with creative behaviour). Existing large scale surveys, such as PISA or TIMMS, are mainly convergent thinking tests. This means that in all the items proposed there is only one correct answer. From the description above on creativity, it can be said that trying to measure creativity with items that have been design specifically to test knowledge in one area (mathematics, science or reading) will present several challenges.

First, the differentiation of creative thinking from knowledge proficiency might be difficult to achieve. It could be argue that certain items in the PISA study would require more creativity than others. After all, questions in PISA require non-routine answers to problems that the respondent might have never encountered before. Thus, if it were possible to differentiate between more or less creative items in the PISA test, it would be possible to create a sub-scale of "creativity" using PISA. In this way, we would have a rough level of 15-years old creativity level for comparison.

This type of measurement, however, opens an important question: how could we say that those items have been correctly answered by the more creative student and not by the most knowledgeable? The items have been constructed not to measure creativity (can the students provide an original and adequate answer?) but to measure their proficiency level in one area (can the student provide one adequate answer?). Thus it might not be possible realistically to separate the "proficiency" to the "creative" part of solving a problem in PISA.

How to decide which items require more creativity than others might be complicated. A group of experts could explore items and decide on this issue, but the items selected as "creative", would have been selected more or less arbitrarily (by a group of experts). Very likely, the resulting scales would have the low reliability, which will make very complicated any interpretation of the results. We will be measuring, not only a small fraction of what apparently (and according to some 30 or 20 experts) creativity entitles, but we will not be able to be sure that we are actually measuring what we are suppose to measure (e.g. creativity in solving mathematical problems). How could we overcome such an obstacle? Can we empirically find evidence that some items are more creative than others?

Another important issue that appears when measuring creativity refers to the conditions of the test. When revising evidence from Wallach and Kogan (1965), Runco (2007, 3) noted: *"if schools care about creativity and give children exercises and test for creativity potential, but if those are given in test-like academic atmosphere, the same children who always do well on test will excel, and the children who do moderately or poorly on traditional tests will again do only moderately or poorly."*

The application of the creativity tests has to be in a "game like" or "permissive environment". Only if tests were described as games rather than tests, where no grades would be given, results would be significantly different than a test of intelligence.

More complicated is to think of creativity as a social construct, in the sense that creative products are determined by the culture in where they have been produced. To put an extreme example: "using a lever to move a rock might be judge novel in a Cro-Magnon civilization, but not in a modern one" (Flahery 2005, 147). In this way, creativity would be context dependent and difficult to be compared across countries. It will be important to determine what aspects of creativity could be measured that are comparable across cultures.

Literature on creativity presents a rather complicated view of what creativity is. The most advanced theories on creativity present a multidisciplinary approach, in which creative behaviour and thinking emerged from the combination of the right variety and level of elements in a very sophisticated interrelationship. Thus, existing international surveys could only provide a (very small) fraction of what creativity encompassed.

Developing a large scale survey on creativity - Is CBA the answer?

From the arguments presented above, it seems costly and maybe not very effective to use existing international surveys as a measure of creativity. The development of an instrument to test creativity in an international comparative way would be an extremely challenging process. However, new methods of assessment with computers might allow for a more efficient way of testing complex constructs such as creativity. Could the emergence of Computer-based Assessment open a door to make the measurement of creativity at an individual level plausible in an international comparative manner?

The first step would necessarily be to create a working definition of creativity. The definition would require the participation of as many stake holders as possible, in order to make it relevant to different policy areas. It would have to be adaptable to many different backgrounds and cultures that will be tested. The definition of creativity would delineate what type of creativity in terms of "originality" we are interested in. As indicated above, originality refers to the fact that something can be "new" to an individual, to a reference group or to the whole society. The later refers to eminent-level creativity and it is only possible to be measured a posterior. Reference group originality, would necessarily mean that we have to define the reference group. Using, for example, experts in some area to assess the "creativity" of a given answer would fall under this category. Individual level creativity means that something is new to the person. This is the every-day-creativity, where a person is able to provide a new and adequate answer to a problem.

In addition to deciding the originality level, it would be necessary to determine if we are thinking of creativity as a characteristic of people, processes or products. Depending on the choice, a variety of testing methods would have to be used. In each of the cases we would have to determine what are the characteristics of creativity, that make it different from a non or less creative person, process or product. This requires the creation of a creativity framework that would list the necessary features that encompass a creative person, process or product.

Once there is a working (agreed) definition of creativity, and resolve (or delineate) some of the main problems facing the construct, it would be advisable to conduct a feasibility study. The objective of this feasibility study would be to determine what aspects of this definition are subject to measurement. This will require some expert group in measurement issues (and creativity research specifically) that would assess the possibility of developing a test to measure creativity in an international manner. If the feasibility study yields a positive result (that is to say, experts in the area agree that something can be measured in an international comparative way) it would provide the "green light" to start the process of developing a tool to measure creativity. The instrument would have to be tested and adapted to the national contexts. The tool would have to necessarily be pre-tested in as many different countries as possible. If the results of these pilot tests are satisfactory, and the tool is good enough, it would be possible to start a full scale process that would provide a picture of creative levels in young people.

The instrument would most likely have to be computer-based. One advantage of computer-based assessment is that it would be able to address some of the creativity measurement challenges better than traditional paper and pencil test. For example, through new methods of simulation, it is possible (at least partially) to monitor the process that a person follows to arrive to a solution. This would allow to test for creative aspects of the process that are not possible in traditional paper and pencil tests. Also important is the fact that computer-based assessment can be easily presented as a game-like test that seems necessary to properly test creativity (Runco 2007).

The project of creating a measurement tool for creativity at the individual level would necessarily be a long term project that requires an important amount of investment and political will. Great difficulties could be expected in the measurement and adaptation of test items to specific cultural aspects. The results of such a project are difficult to predict, as well as the reactions that might cause in the public sphere.

References

Amabile, T. M. (1983). The social Psychology of creativity. New York. Springer-Verlag.

Andreasen, N. (2005). The creating brain. New York: Dana Press.

Bailin, A. (1988). Achieving extraordinary ends: An essay on creativity. Boston, MA: Kluwer Academic.

Ball, O. E. and Torrance, E. P. (1980). Streamlined Scoring and Interpretation Guide and Norms for the Figural Torrance Test of Creative Thinking, Form A. Arhens, GA: Georgia Studies of Creative Thinking (mimeo).

Bean, R. (1992). How to develop your children's creativity. Los Angeles, CA: Price stern Sloan Adult.

Csikszentmihalyi (1996). Creativity. New York: HarperCollins.

DCMS (Department for Culture, Media and Sport) (2006). Government Response to Paul Robert's Report on Nurturing Creativity in Young People. London: DfES.

European Commission (2008). Proposal for a Decision of the European Parliament and of the Council concerning the European Year of Creativity and Innovation (2009). COM (2008) 159, final. Brussels: European Commission.

Flaherty, A. W. (2005) Frontotemporal and dopaminergic control if idea generation and creative drive, Journal of Comparative Neurology, 493, 147-153.

Florida, R. (2002) The rise of the creative class... And how it's transforming work, leisure, community and everyday life. New York: Basic Books.

Florida, R. (2004) The rise of the creative class... And how it's transforming work, leisure, community and everyday life (paper back). New York: Basic Books.

Getzels, J. W. and Jackson P. W. (1962). Creativity and intelligence: Explorations with gifted students. New York: Wiley.
Gough, H. G. (1952). Adjective Check List. Palo Alto: Consulting Psychology press.

Gruber, H. E. and Davis, S. N. (1988). Inching our way up Mount Olympus: The evolving-system approach to creative thinking. In R. J. Sternberg (ed.) The nature of Creativity (pp. 243-270). New York: Cambridge University Press.

Gruber, H. E. and Wallace, D. B. (1999). The case study method and evolving system approach for understanding unique creative people at work. In R.J. Sternberg (ed.) Handbook of Creativity, pp. 3-16. London: Cambridge University Press.

Guilford, J. P. (1962). Creativity: Its measurement and development. In J. J. Parnes and H. F. Harding (eds.), A source book for creative thinking. New York: Scribners.

Haensly, P. A., and Torrance, E. P. (1990). Assessment of creativity in children and adolescents. In Reynoolds, C. R. and Kamphaus, R. W. (eds), Handbook of Pshycological and Educational Assessment of Children: Intelligence and Achievement.

Hattie, J. A. (1977). Conditions for administering creativity tests, Psychological Bulletin, 84, 1249-1260.

Heausler, T .P., and Thompson, B. (1988). Structure of the Torrance Tests of creative thinking, Educational and Psychological Measurement, 48, 463-468.

Herbert, T. P., Cramond, B., Neumeister, K. L. S., Millar, G., and silvian, A. F. (2002). E. PaulTorrance: His life, accomplishments, and legacy. Storrs: University of Connecticut, The National Research Center on Gifted and Talented (NRC/GT).

Hocevar, D. (1981). Measurement of Creativity: Review and Critique, Journal of Personality Assessment, 45, 450-464.

Houtz, J. C. and Krug, D. (1995). Assessment of Creativity: Resolving a Mid-Life Crisis, Educational Psychology Review, 7 (3), 269-300.

KEA European Affairs (2006). The Economy of Culture in Europe. Brussels: European Commission, DG Education and Culture.

Kim, K. H. (2006). Can we trust creativity tests? A review of the Torrance Tests of creative thinking (TTCT), Creativity Research Journal, 18 (1), 3-14.

Lissitz, R. W. and Willhof, J. L. (1985). A methodological study of the Torrance Tests of Creativity, Journal of Educational measurement, 22, 1-111.

Nonaka, I. (1991). The knowledge creating company, Harvard Business Review, 69, pp. 96-104.

Nonaka, I. and Takeuchi, H. (1995). The knowledge creating company: How Japanese companies create the dynamics of innovation. Oxford: Oxford University Press.

Mayer, R. E. (1999). Fifty years of Creativity Research. In R.J. Sternberg (ed.) Handbook of Creativity, pp. 449-460. London: Cambridge University Press.

McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience, Journal of Personality and Social Psychology, 52, 1258-1265.

Mednick, S. A. (1962). The associative basis for creative process, Psychological Bulletin 69, 220-232.

Mumford, M. D. (2003). Where have we been, where are we going? Taking stock in creativity research, Creativity Research Journal 15, 107-120.

NACCCE (National Advisory Committee on Creative and Cultural Education) (1999). All Our Futures: Creativity, Culture and Education. London: DfES.

Plass, H., Michael, J. J., and Michael, W. B. (1974). The factorial validity of the Torrance Test of Creative Thinking for a sample of 111 sixth-grade children, Educational and Psychological Measurement, 34, 413-414.

Plsek, P. E. (1997). Creativity, Innovation and Quality. Quality Press.

Plucker, J. A. and Renzulli, J. S. (1999). Psychometric approaches to the Study of Human Creativity. In R.J. Sternberg (ed.) Handbook of Creativity, pp. 35-62. London: Cambridge University Press.

Richards, R. (1999a). Everyday Creativity. In M. A. Runco and S. Pritzker (eds), Encyclopedia of Creativity, 683-689. San Diego, CA: Academic Press.

Richards, R. (1999b). The subtle attraction: Beauty as the force in awareness, creativity, and survival. In S. W. Russ (Ed.), Affect, creative experience, and psychological adjustment, 195-219. Philadelphia: Brunner/Mazel.

Rodhes, M. (1962). An analysis of creativity, Phi Delta Kappan 42, 305-310.

Rosenthal, A., DeMers, S. T. Stiwell, W., Graybeal, S., and Zins, J. (1983). Comparison of interrater reliability on the Torrance Test of Creative Thinking for gifted and nongifted students, Psychology in the Schools, 20, 35-40.

Runco, M. A. (2003). Discretion is the better part of creativity: Personal Creativity and Implications for Culture, Inquiry: Critical Thinking Across the Disciplines 22, 9-12.

Runco, M. A. (2004). Personal creativity and culture. In S. Lau, A. N. N. Hui and G. Y. C. Ng (eds), Creativity when East meets West, 9-22. New Jersey: World Scientific.

Runco, M. A. (2007). Creativity. Theories and Themes: Research, Development and Practice. Amsterdam: Elsevier.

Runco, M. A., and Albert, R. S. (1986). The threshold hypothesis regarding creativity and intelligence: An empirical test with gifted and nongifted children, Creative Child and Adult Quarterly, 11, 212-218.

Simonton, D. K. (1990a). In M. A. Runco and R. S. Albertr (eds), Theories of creativity. Newbury Park, CA: Sage.

Simonton, D. K. (1999). Creativity from a Historimetric Perspective. In R.J. Sternberg (ed.) Handbook of Creativity, pp. 117-133. London: Cambridge University Press.

Smith, S. M., Ward, T. B., and Finke, R. A. (1995) (eds). The creative cognition approach. Cambidge University Press.

Solomon, B., Powell, K., and Gardner, H. (1999). Multiple Iintelligences. In M. A. Runco and S. Pritzker (eds), Encyclopedia of creativity,259-273. San Diego,Acedemic Press

Sternberg R. J. and Lubart, T. I. (1992). Buy low and sell high: An investment approach to creativity, Current Directions in Psychological Science, 1 (1), 1-5.

Sternberg R. J. and Lubart, T. I. (1995). Defying the crowd: Cultivating creativity ina culture of conformity. New York: Free Press.

Sternberg R. J. and Lubart, T. I. (1996). Investing in creativity, American Psycgologist, 51, 677-688.

Sternberg, R.J. and Lubart, T. I. (1991).An investment theory of creativity and its development,Human Development,34,1-32.

Sternberg, R. J. (1999). A propulsion theory of creative contribution, Review of General Psychology, 3, 83-100.

Sternberg, R. J. (2006a). Introduction. In J. C. Kaufman, R. J. Sternber (eds). The International Handbook of Creativity, 1-10.

Sternberg, R. J. (2006b). The nature of creativity, Creativity Research Journal, 18 (1) 87-98.

Sternberg, R. J. and Lubart, T. I. (1999) The concept of creativity: Prospects and Paradigms. In R.J. Sternberg (ed.) Handbook of Creativity, pp. 3-16. London: Cambridge University Press.

Torrance, E. P. (1966). The Torrance Tests of Creative Thinking-Norms-Technical Manual Research Edition-Verbal Tests, Forms A and B-Figural Tests, Forms A and B. Princeton, NJ: Personnel Press.

Torrance, E. P. (1974). Torrance Test of Creative Thinking. Lexington, MA: Personnel Press., p. 6).

Villalba, E. (2008). The Uniqueness of Knowledge Management in Small Enterprises: Managing Knowledge as an Employer Strategy for Lifelong Learning. Saarbrücken: VDM Verlag.

Wallach, M. A. and Kogan, N. (1965). Modes of Thinking in Your Children: A Study of the Creativity-intelligence Distinction. New York: Holt, Rinerhart & Winston.

Wehner, L., Csikszentmihalyi, M. and Magyari-Beck, I. (1991). Current approaches used in studying creativity: An exploratory investigation, Creativity Research Journal 4, 3, p. 261-271.

The author:

Ph.D. Ernesto Villalba
European Commission- Joint Research Centre, IPSC
Centre for Research on Lifelong Learning (CRELL)
Via Fermi, 2749, 21027 Ispra (VA), Italy
Tel: +39-0332-785226
E-Mail: ernesto.villalba@jrc.it

Ernesto Villalba is a Scientific Officer Centre for Research on Lifelong Learning (CRELL) at the Joint Research Centre (JRC) of the European Commission. He is leading the project “education for innovation and innovation for education” dealing with the relationship between education, creativity and innovation. He holds a Ph.D. in international and comparative education from the Institute of International Education at Stockholm University. His main areas of interest are: lifelong learning, innovation, creativity and knowledge management.

Experiences from Large-Scale Computer-Based Testing in the USA

Brent Bridgeman
Educational Testing Service, USA

Abstract

Computer-based tests offer numerous advantages over paper-based tests. Advantages include: paperless test distribution and data collection, greater standardization of test administrations, monitoring of student motivation, obtaining machine-scorable responses for writing and speaking, providing standardized tools for examinees (e.g., calculators and dictionaries), and the opportunity for more interactive question types. Each of these advantages comes with a series of challenges that must be addressed if the computer-based test is to be effective and fair. Research has not only identified some of these potential problems, but has also suggested solutions to many of them.

Computer-based testing is now common, but not ubiquitous, in the United States. The major college admissions tests, the SAT and ACT, still rely on paper-based tests because of the need to test very large numbers (in the millions) in a couple of months. But other large-scale testing programs based in the US have embraced computer-based tests (CBTs). Some of these are computer-adaptive tests (CATs) in which examinees are branched to easier or harder questions based on their performance on prior questions, and others use computers to administer and score non-branching linear tests. Some of the major computer-based tests in the US are:

- Armed Services Vocational Aptitude Battery (CAT-ASVAB)—this test is administered to high school students, especially those interested in pursuing careers in the military, and helps to identify the military specialties for which they would be best suited.
- Measures of Academic Progress—a battery of CATs that are used for K-12 assessments in reading, mathematics, and language usage in over 3.400 school school districts.
- US Medical Licensing Examinations—a series of three examinations that MD candidates take as they progress through medical school and into residency programs.
- Microsoft Certification Examinations—a mixture of linear CBTs, CATs, and computer simulations for certifying candidates to work on computer software problems.
- Graduate Management Admissions Test—a CAT used as part of the admissions process for graduate management in the US and in English-speaking programs abroad.
- Graduate Record Examination-General Test (GRE CAT)—a CAT used as part of the admissions process to a variety of graduate programs at the masters and doctoral levels.
- The Praxis Series: Teacher Licensure and Certification, Praxis I, Pre-Professional Skills Test—linear CBT testing basic skills in reading, writing, and mathematics
- National Council of Architectural Registration Boards—used as part of the process for licensing architects, this test consists of linear multiple-choice CBTs in six areas plus three graphic problems answered on screen and scored by computer that cover aspects of site planning, building planning, and building technology.
- Test of English as a Foreign Language, Internet-Based Test (TOEFL iBT)--linear CBT that assesses ability to read, listen, write, and speak in English and is used for admission to undergraduate and graduate programs in the US and Canada.
- Test of English for International Communication (TOEIC)--linear speaking and writing tests delivered by computer and, in combination with some paper-and-pencil reading and listening assessments, used by businesses that need an assessment of English skills of potential employees.

This listing is intended simply to give an overview of the broad variety of CBTs (including but not limited to CATs) administered in the US (and in many cases internationally by US companies); it is certainly not a complete catalogue of computerized tests. What this list should imply is that there are numerous advantages to CBTs that have been attractive to these testing programs for different reasons, but each of these programs also has had to overcome challenges in test development and test delivery to make the CBTs efficient, effective, and fair. This paper presents some of the advantages of switching from paper-based testing to CBTs, but notes that with each of these advantages there are also challenges that must be addressed and overcome. Research has identified many of these challenges, and in some cases suggested approaches to overcoming the problems.

CBT Advantages—Paperless Test Distribution and Data Collection

Printing test booklets and mailing large quantities of them to test centers can be a major expense for testing programs. If an error is found after booklets are printed (often months in advance of the actual testing) booklets must be reprinted and reshipped at considerable expense. If the test is simply an electronic file, it can be relatively easily corrected at any point prior to test administration, and it can then be sent electronically over the Internet to testing locations all over the globe for very little expense. Electronic delivery also provides substantial advantages for test security. Instead of test booklets sitting in offices for days or even weeks before a test administration (with the opportunity for a booklet to be stolen and distributed before the test), tests can be sent over the Internet at the last minute, or even while the test is in progress, thus reducing the possibility of questions being exposed prior to the test.

After the test, there is no need to mail answer sheets back to a central location for scoring with a chance that they could be lost in the mail. Computer delivery also allows for the possibility of instant scoring. With the GRE CAT, for example, examinees can get a preliminary view of their scores immediately after answering the last question.

With computerized data collection, different kinds of data can be obtained. For example, data on the amount of time spent on each question can be easily collected. Also, data can be collected for alternative response formats, such as answer until correct.

CBT Challenges—Paperless Distribution and Data Collection

Despite the many advantages of paperless test distribution and data collection, there are also potential problems and challenges. A momentary power interruption has no impact on delivery of a paper test, but can make it necessary to reboot computers. Although recovery systems can be built so that data is not lost in the reboot, and examinees can restart the test in the same spot, building and testing such systems can be expensive. Use of laptops or other systems with battery backup can also address the momentary power failure problem.

Computer delivery requires attention to hardware issues that are not relevant for paper-based tests. For example, screen displays must be standardized for all examinees. Bridgeman, Lennon, and Jackenthal (2003) showed that examinees who took a reading test on a large, high-resolution monitor scored significantly higher than examinees who took the test on a smaller low-resolution monitor. It appeared that the major problem was that texts on a low-resolution screen required scrolling to see the entire passage while on a high resolution screen the entire passage was visible while questions were being answered. Differences in keyboarding skills may also become important if the test requires examinees to write an essay. Bridgeman and Cooper (1998) administered GMAT essays in both paper-and-pencil and word-processed formats to the same people and computed the difference between the scores in the two formats. They found that this difference was substantially larger for examinees with relatively little word processing experience (less than once a week) compared to examinees who reported using a word processor more than two times a week. Similarly, a study of essays written by 8th grade students as part of a pilot project for the National Assessment of Educational Progress found that scores on handwritten essays could differ from scores on a computer-based test for students with relatively low levels of computer familiarity; specifically, the computer familiarity predicted online writing score after controlling for paper writing score (Horkay, Bennett, Allen, Kaplan, & Yan, 2006). The type of keyboard may also be important. Powers and Potenza (1996) compared GRE tests given on a laptop computer or on a desktop computer. For multiple-choice questions the computer type did not matter, but for essays, scores were generally lower for the examinees who had to write on a laptop. But the date on this study may be significant; in 1996 relatively few students used laptops, but now they are the standard computer type for many if not most university students. It is conceivable that if this study were replicated today, it might be the desktop users who were at a disadvantage. The most likely result is that it depends on the computer type that is most familiar to the examinee, but this creates a dilemma for a standardized testing situation where it might not be possible to match students with their preferred computer type.

CBT Advantages—Greater Standardization of Test Administrations

Standardized testing requires test administrations to be as nearly equal as possible for all examinees. If time is at all a factor in a test, it is imperative that the timing conditions be the same for all examinees. If the test administration relies on human examiners watching a clock, sometimes some examinees are inadvertently given more time than others. Computers can manage test timing very accurately, assuring fair timing for all. At the individual item level, the computer can accurately record reaction time for simple prompts or solution times for more complex problems.

CBT Challenges—Greater Standardization of Test Administrations

Although computers can be very accurate time keepers, tests with strict time limits can be very problematic for certain kinds of computerized tests. Specifically, CATs with strict time limits can raise substantial fairness concerns. If the item selection algorithm is based only on item difficulty and discrimination, and not on the amount of time required to answer a particular question, it is possible for some examinees to get tests that are more time consuming than others. Bridgeman and Cline (2000) showed that math items at the same difficulty level can vary greatly in the amount of time required to answer them. Furthermore, because the CAT scoring algorithm assumes that an incorrect answer implies low ability (not high ability students running out of time and guessing randomly), a series of incorrect answers by students guessing as time runs out can dramatically lower scores by as much as two standard deviations over what the score would have been before the random guessing began (Bridgeman & Cline, 2004). Although models have been developed to describe the impact of time limits on CAT scores (e.g., Schnipke & Scrams, 1997; van der Linden, 2008), there is no adequate way to assess what scores would have been without the time limit.

CBT Advantages—Monitor Student Motivation

National educational surveys, such as the National Assessment of Educational Progress in the United Sates, or international surveys, such as TIMMS and PISA, must assume that students taking the test are making an honest effort to answer the questions to the best of their abilities. But there are no consequences to the individual test takers for poor performance. If results from these surveys are to be believed, ways to screen out responses from unmotivated test takers must be found. Fortunately, computer delivery can provide some useful tools for identifying these students. In particular, by monitoring the time spent on each question, the computer can identify students who are responding at an unreasonably fast rate that suggests they are not fully considering the question. A slightly more sophisticated approach could monitor differences in the time spent on questions that most students can answer quickly in comparison with questions that should take more time; students who take an equal amount of time for both types of questions may not be seriously considering each question. Responses from apparently unmotivated students can be documented and removed from the analysis, but an even more effective approach may be to use real-time monitoring to motivate these students. Students who are informed that the computer has identified them as not trying their best can be encouraged to do better. Students who are monitored and encouraged to do better actually do improve their performance with positive impacts on test validity (Wise, Wise, & Bhola, 2006). This kind of motivational monitoring is also useful in experimental studies of new item types or other test features in which conclusions should be based only on responses from student who are making an honest effort.

CBT Challenges—Monitor Student Motivation

There appear to be few problems with motivation monitoring, but there might be privacy concerns if item timing information is being collected without the students' knowledge or consent. If students are informed about the monitoring, this could increase anxiety levels. Although increasing anxiety levels for unmotivated students may actually lead to more valid scores, increasing anxiety for students who were already trying hard could lead to poorer performance.

CBT Advantages—Obtain machine-scorable responses for writing and speaking

If written responses are entered on a computer, they can be electronically scored, resulting in substantial savings in payments to raters. Computerized natural language processing tools can be used to assess essay features such as organization, development, grammar, and mechanics. Electronic scoring of essays closely mimics the results of human scoring, and the agreement of an electronic score with a human score is typically as high as the agreement between two humans, and sometimes even higher (Attali & Burstein, 2006). Similarly, spoken responses can be captured by the computer and automatically scored. Automated scoring for highly predictable speech, such as a one sentence answer to a simple question, correlates very highly with human ratings of speech quality. For longer and more open-ended responses, automated speech scoring is not yet good enough for use in high stakes tests, but the technology is evolving rapidly and appears to be adequate for lower stakes practice tests (Xi et al., 2008).

CBT Challenges—Obtain machine-scorable responses for writing and speaking

Although machine scoring works well on average, the machine cannot evaluate the quality of an argument. A long, grammatical essay may receive a high score from a machine even if the argument is fallacious. Examinees who know in general terms how the machine scoring works may be able to produce essays that will fool the machine into giving an essay a higher score than it deserves, although fooling the system is not as easy as some might believe (Powers et al., 2002). Because it is possible to fool the machine with well-written nonsense, most high stakes tests that use a machine in some capacity also have all essays read by at least one human. An additional problem is that the public may not accept an essay score provided by a machine regardless of what research says about the validity of the scores. One approach to this problem is to have the machine act just as a quality control device that will flag essays for which the human and machine scores are discrepant. The flagged essays are then given to a second human. The flagging improves the reliability of the scoring, but a machine score is never part of the score assigned to an essay (Monaghan & Bridgeman, 2005).

CBT Advantages—Additional Tools for Examinees (e.g., calculators, dictionaries)

If a test is designed to assess mathematical reasoning abilities, it may be desirable to minimize simple computational errors by providing a calculator for the test. Furthermore, students routinely use calculators for classroom activities and homework, so it would be reasonable to also provide a calculator for the test. With a computer-delivered test, the calculator can be provided as an onscreen tool--the same tool for all examinees. The calculator can be tailored to the needs of a particular test, so one test might require a full scientific calculator while another test would require only a simple four-function calculator. In addition, the calculator can be turned on for certain items and turned off for others; the calculator could interfere with the assessment for test items designed to assess computational skill or estimation skill. Although hand-held calculators can be provided for paper-based tests, this ability to make calculator use item specific is possible only with a computer-delivered test. Other tools, such as dictionaries, can also be provided on an as-needed basis. Dictionaries may be especially useful for tests of non-native speakers who have difficulty with an item just because of an unfamiliar word. As with a calculator, it may be important to be able to turn off the dictionary for certain questions, if, for example, the item is intended as a vocabulary measure for a specific word.

CBT Challenges—Additional Tools for Examinees (e.g., calculators, dictionaries)

Although providing one common on-screen calculator for all examinees has advantages, it also creates some fairness concerns. For some examinees, the common calculator may function in a manner that is different from the one that they usually use. For example, some calculators respect order of algebraic operations and others do not, so in some calculators 5+2x3 will equal 11 while in other it will equal 21. Therefore, it is essential to provide pre-test practice with the calculator that will be used during the test.

CBT Advantages—More Interactive Question Types

Although many first-generation CBTs simply administered standard multiple-choice questions via a computer, the next generation may make much more extensive use of the computer's capability to administer items that go far beyond multiple-choice. For example, examinees can be asked to perform on-line experiments, make data plots, and then answer questions related to the experiment they just performed. Experimental items using this approach were successfully pilot tested for the U. S. National Assessment of Educational Progress (Bennett, Persky, Weiss, & Jenkins, 2007). One test has used very sophisticated computer-delivered and computer-scored graphical problems since 1997. In the licensing test for architects for the National Council of Architectural Registration Boards, the examinee completes design problems on screen (e.g., designing a portion of a building such that requirements related to fire codes and access for disabled persons are met), and the machine automatically evaluates the adequacy of the solution (Bejar & Braun, 1999).

CBT Challenges—More Interactive Question Types

Designing clever interactive questions is sometimes easier than designing effective scoring strategies for these items (Bennett, 2006). Assigning an appropriate score for an interactive design problem with open-ended responses is much more difficult than developing an answer key for a multiple-choice question. Furthermore, questions that appear to be unique and innovative may be tapping the same construct as far simpler and cheaper multiple-choice counterparts. Questions on the GRE analytical reasoning test specified a series of conditions and then asked a series of multiple-choice questions. For example:

> *An organist is arranging to judge the playing of original compositions by six*
> *student organists-- R, S, T, U, V, and W. She will hear one student play each*
> *day from Monday through Saturday. She must schedule the auditions for the*
> *students according to the following conditions:*
> *R must play earlier in the week than W.*
> *S must play on Thursday.*
> *T must play on the day immediately before or immediately after the day*
> *on which U plays.*
> *V cannot play on Tuesday.*
> *The organist could schedule any of the following to play on a day*
> *immediately before or after the day on which T plays EXCEPT*
> *(A) R (B) S (C) U (D) V (E) W*

An experimental computer-delivered version of this item was created that provided the examinee with a calendar and asked the examinee to place the student organists in the calendar in such a way that additional constraints were met. For example,

If R must play on the day immediately after the day on which V plays, make a possible schedule of auditions.

The computer could handle multiple correct answers, and in this case there were two: VRWSTU or VRWSUT. Placing the six student organists into the schedule with the possibility of more than one correct answer seemed to be a much more natural task than answering a multiple-choice question, but the underlying skill was really the same. In order to answer the multiple-choice questions, the examinee still had to create the calendar on scratch paper. Factor analyses confirmed that both formats seemed to be tapping the same skill (Bridgeman & Rock, 1993). Although there is some value in just improving the face validity of an assessment instrument, it may be difficult to justify the additional costs if there is no measurable gain in validity.

Conclusion

Computerized testing provides many benefits over paper-based tests. Operationally, handling electronic files has advantages relative to paper, and there are monitoring possibilities, additional tools, and innovative item types that are simply not possible with paper-based tests. But before rushing to embrace computer-based tests, users should also be aware that each advantage of this new technology also provides challenges that may not be immediately apparent.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. Journal of Technology, Learning, and Assessment, 4(3). Retrieved Nov. 13, 2008 from http://www.jtla.org.

Bejar, I. I., & Braun, H. I. (1999). Architectural simulations: From research to implementation. (ETS-RM–99–2). Princeton, NJ: Educational Testing Service.

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, NJ: Erlbaum.

Bennett, R.E., Persky, H., Weiss, A.R., and Jenkins, F. (2007). Problem Solving in Technology-Rich Environments: A Report From the NAEP Technology-Based Assessment Project (NCES 2007–466). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved Nov. 12, 2008 from http://nces.ed.gov/nationsreportcard/pdf/studies/2007466.pdf

Bridgeman, B. & Cline, F. (2000). Variations in Mean Response Times for Questions on the Computer- Adaptive GRE General Test: Implications for Fair Assessment. (Graduate Record Examination Board 96-20P; ETS RR-00-07). Princeton, NJ: Educational Testing Service.

Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. Journal of Educational Measurement, 41, 137-148

Bridgeman, B., & Cooper, P. (April, 1998). Comparability of scores on word-processed and handwritten essays on the Graduate Management Admissions Test). Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Bridgeman, B., Lennon, M., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. Applied Measurement in Education, 26, 191-205.

Bridgeman, B., & Rock, D. (1993). Relationships among multiple-choice and open-ended analytical questions. Journal of Educational Measurement, 30, 313-329.

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). Retrieved Nov. 13, 2008 from http://escholarship.bc.edu/jtla/ vol5/2/.

Monaghan, W, & Bridgeman, B. (2005). E-rater as a quality control on human scores. (ETS-RDC-02). Princeton, NJ: Educational Testing Service.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. Computers in Human Behavior, 18,103-134.

Powers, D., & Potenza, M. (1996). Comparability of testing using laptop and desktop computers. (ETS RR-96-15). Princeton, NJ: Educational Testing Service.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. Journal of Educational Measurement, 34, 213-232.

van der Linden, W. (2008). Using response times for item selection in adaptive testing. Journal of Educational and Behavioral Statistics, 33, 5-20.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. Educational Assessment, 11, 65-83.

Xi, X., Higgins, D., Zechner, K., & Williamson, D.M. (2008). Automated scoring of spontaneous speech using SpeechRater v1.0. Princeton, NJ: Educational Testing Service.

The author:

Brent Bridgeman
Educational Testing Service, 09-R
Princeton, NJ 08534, USA
E-Mail: bbridgeman@ets.org

Dr. Bridgeman's title is Principal Research Scientist in the Center for Validity Research at Educational Testing Service. He is interested in a broad range of validity issues, especially threats to validity from construct underrepresentation and construct irrelevant variance.

National Tests in Denmark – CAT as a Pedagogic Tool

Jakob Wandall
Skolestyrelsen, Denmark

Abstract
Testing and test results can be used in different ways. They can be used for regulation and control, but they can also be an important pedagogic tool to assess student proficiency in order to target teaching and improve learning. This article gives the background and status for the development of the Danish national tests. It describes what is special about these tests (IT-based, 3 tests in 1, adaptive, etc.), how they are carried out and what is tested. The national test are supposed to be low stake, and to secure this, the results are confidential. It is described who are allowed to know the results, what kind of response is given to the pupil, the parents, the teacher, the headmaster and the municipality and how the teacher and headmaster can use the results. The only results that are made public are the overall national results. Because of the test design (Rasch-model) test results can be compare without any restrictions, which gives an enormous potential for developing new ways of using test results in a pedagogical context.

Background and status for the development of the Danish national tests

On April 19th, 2006 the Danish Parliament decided to make national tests a compulsory pedagogic tool in the *Folkeskole*[1]. The tests are part of the implementation of the recommendations from a review conducted in Denmark by a team from OECD. The tests are intended to support improvement of the evaluation culture in Denmark.

The tests are designed by *The Agency for the Evaluation and Quality development of Primary and lower Secondary education* and are developed by a consortium (COWI A/S in cooperation with different companies as well as educational and research institutions).

The development of the IT-based test system began in July 2006, and the first three tests were launched in May 2007. This first version was reviewed by an expert panel. This group of experts concluded that the basic concept was very successful – even though there were some teething troubles. But some more serious problems were also detected: The test items simply did not have the sufficient quality, neither was the quantity of items sufficient. Therefore, it was necessary to redesign and try out all the items and to enlarge the item banks. This process should, according to the plan, be finalised by the end of 2008.

Even though the work has been going on for 2½ years, there are still some questions yet to be answered. The next version of the tests will be launched as a pilot version in the spring 2009. The first three years can be considered as a period of learning, and a lot of changes and improvements have been made according to the experiences achieved during this period of time.

What is special about these tests?

The national tests differ from the tests that are already used in the Danish schools in the following ways:

- The tests are IT-based and the pupils answer the questions online.
- Test results (scores and reports of results) are automatically calculated and generated. The teachers do not have to correct the tests, and the analysis of test results (or some parts of the work) has been done when the teacher gets the results.
- The *Agency for the Evaluation and Quality development of Primary and lower Secondary education* supplies the schools with the tests free of charge.
- The tests are adaptive. Each test contains three separate adaptive test sessions which deal with different dimensions of the subject (so-called “profile areas”, described below).

Adaptive tests adapt to the pupils’ level of proficiency during the test. The first item presented to the pupil has an average difficulty (compared to the form the test is designed for). If the answer is correct, the next item presented to the pupil will be more difficult. If the answer is wrong, the next item will be easier. In this way the test will adapt to the pupils’ level, so that the sequence of items will be different for each pupil. This is a very simplified description of the principle that is employed during the entire test.

[1] *Folkeskolen* is the Danish term for the public Primary and lower Secondary School, more info in English, see: http://eng.uvm.dk/~/media/Files/English/Fact%20sheets/2008_fact_sheet_the_folkeskole.ashx

Item difficulty and pupil ability

In order to match the item difficulty with the pupil's ability, they have to be measured on the same scale. Several item response models were considered, but we chose the original one-dimensional model given by the Danish statistician George Rasch: It is a simple model to use and to handle and fits the purpose (adaptive testing) best. But there is a catch – the Rasch model is very inflexible and usually it leads to scrapping a large proportion of the items.

The item difficulty on the Rasch-scale (also called "theta-scale") is defined as the ability of the pupil which has exactly 50% probability to give the right answer on the item.

In a well-designed ordinary test (linear test, where the series of items is predetermined) most pupils will experience that some items are too easy, others too difficult and (hopefully) some items difficulty fit the individual pupil's ability. From an analytical point of view the test items that are too easy or too difficult, reveal very little about the pupil's ability.

Only the items where the level of difficulty fits the pupil's ability contribute substantially to the estimation of the pupil's ability. And in a well designed adaptive test the pupils will mostly be presented for items that have a suitable level of difficulty for his/her level of ability.

How are the tests carried out?

The test system is connected with the Danish website *evaluering.uvm.dk* – it is through this site both the teacher and the pupils access the test. The test system has a maximum capacity of 6.000 users (pupils) at the same time.

The teacher logs on and opens the access to the testing system for his pupils. Every teacher, headmaster and pupil in Denmark has a unique user-id/password. Like the test system, the identification system is provided by the Ministry of Education. When the pupils are allowed to start the test, they log in.
They now have 45 minutes to answer as many items as possible. During this time, the pupils will typically answer 50-80 questions. If a pupil needs more time, it is possible for the teacher to prolong the test for the individual pupil.

For this purpose, the teacher has a monitor screen that shows the SEM-level (SEM= "Standard Error of Measurement", the statistical validity of the estimate of ability) of test result continuously during the test. It is done with a colour indication: Red means that the pupil has answered less than 5 questions, and therefore, there is no basis for estimation of the student ability. When the pupil has answered 5 questions, the indicator turns yellow, and the student ability is estimated for the first time. Hereafter, the pupil ability is re-estimated for every item answered.

A central computer registers which items have been answered correct, and which items have been answered wrong. For every item the pupil has answered, the student ability will automatically be estimated. According to this estimate, the central computer will choose the next item for the pupil, so that the item difficulty matches the last estimate of the pupil ability as precisely as possible. Then again the same procedure is repeated. When the SEM reaches 0,3 the indicator on the teacher's monitor screen turns green. The principle is illustrated below.

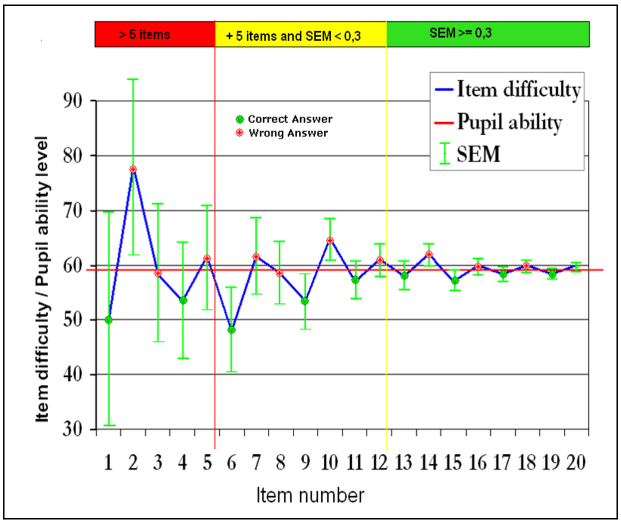


Figure 1: The adaptive principle

Usually SEM-level of 0,3 is reached after 7-12 items. Experience shows that this is more than 3 times faster than average in a linear setup.

The pupils continue to get questions until the time is up. The more items, the better and more detailed the analysis of the pupil's proficiency.

The items are chosen from a database with more than 500 items per test. To measure the single items difficulty, verify that the item fits the scale and secure that the items are of high standard, all items are initially tested on 500-700 pupils. The items and the responses from these initial tests are statistically analysed to eliminate the items that doesn't fit the Rasch model. Items that do not meet these very strict psychometric/statistical demands are not accepted and will not be used in the national test.

What is tested – and when?

12 tests are being developed, with 12 different item banks for 7 different subjects: Danish/reading, math, English, geography, biology, physics/chemistry and Danish as second language. The tests are targeted the form where they are compulsory. 10 of the 12 test are compulsory to use one time per pupil in the *Folkeskole*. The two tests in Danish as second language are voluntary to use for the schools. An overview of the test is given below.

Class Subject	1	2	3	4	5	6	7	8	9
Danish/reading		X		X		X		X	
Mathematics			X			X			
English							X		
Geography								X	
Biology								X	
Physics/Chemistry								X	
Danish as second language					X		X		
X: Compulsory , X: Voluntary									

Figure 2: Tests in the *Folkeskole* class

Compulsory testing in mathematics 6'th form, Danish reading 8'th form and physics/chemistry 8'th form were carried out in May/June 2007. When fully implemented, the compulsory testing must be carried out in all testing subjects every year between February 1st and April 30th.

Within this period of time the teacher can decide when to take the test. The testing system contains a flexible booking facility where the teacher can reserve slots for a pupil or a group of pupils, e.g. a class.

In addition to the compulsory use, the schools are offered to use the testing system voluntarily twice per pupil. Either in the form for which the test is compulsory or in the previous or the following form. Booking for voluntary testing is open all year – except when there is compulsory testing. Due to the construction of the test system, two test results for the same pupil are directly comparable, so the test system is designed for measuring progress and added value.
The next step – when fully implemented – could be to merge the item banks in the same subjects (Danish/reading 2, 4, 6 and 8, Math 3 and 6 and Danish as second language 5 and 7). This way the teacher will be able to monitor progress from 1st to 9th form in Danish/reading, 2 to 7 form in math and 4 to 8 form in Danish as second language. Merging the item banks is being prepared, but is not yet decided.

Which parts of the subject are tested?

The tests are designed to include large and important parts of the subject. However, not all parts of a subject are suitable for this kind of testing, e.g. the pupils' ability to express themselves orally or in writing. Furthermore, the teachers are, due to the law, obliged to assess the student progress regularly. In other words, it is necessary to use other kinds of assessment and evaluation. The national tests can only cover a very small part of the total need for evaluation in the *Folkeskole*.

Therefore the website www.evaluering.uvm.dk also contains a description and a user guide to a large number of other evaluation tools.

Like other kinds of evaluation in the *Folkeskole*, the national tests are carried through as an integrated part of the education.
Every subject is divided into 3 dimensions or areas of the subject, the so-called profile areas, to make a more detailed and precise evaluation of the pupil's proficiency possible.

When being tested, the pupil experiences that the questions are being presented in random order. But as previously mentioned, three separate adaptive test sessions are simultaneously conducted, where the selection of the next item in a given profile area solely depends on the pupil's response of the previous items in the same profile area.

Physics/chemistry

- Profile area 1: Energy and transformation of energy
- Profile area 2: Phenomena, matter and materials
- Profile area 3: Physics/chemistry – applications and perspectives

Mathematics

- Profile area 1: Numbers and algebra
- Profile area 2: Geometry
- Profile area 3: Practical use of mathematics

Reading (Danish)

- Profile area 1: Language comprehension
- Profile area 2: Decoding
- Profile area 3: Reading comprehension

Figure 3: The "profile areas" for the first 3 tests

What is at stake?

Testing and test results can be used in different ways. They can be used as tools for the teacher in order to assess the pupil's knowledge and competencies or it can be used as a pedagogic tool (typically low stake) to assess the effect of the teaching. But it can also be used for admission, regulation, controlling, rewarding/punishment of individuals/schools (typically high stake). From an international perspective, educational testing usually is high stake, i.e. tests for which the results have significant consequences for individuals or schools (e.g. pupils' further educational possibilities, teacher's salary or school grants).

Testing and test results can to some degree be used for both purposes, but there are some restraints.

In high stake testing security, equal terms and fair conditions are key issues. But if the main purpose – as it is in Denmark - is to assess student proficiency in order to target teaching and improve learning (a pedagogic tool), the teacher should have access to full control over the testing conditions (e.g. which aids, tools, remedies and assistive technology are allowed during the test) – in fact, if it improves the analysis of the student proficiency, it would make sense that the teacher is allowed to help if e.g. the pupil gets stuck.

In high stake testing the results are usually made public (at least at a school average level) for different purposes (e.g. for school comparison/ranking.). The test should then be designed to measure the educational curriculum, so that "teaching for the test" is no problem.

In low stake testing – as in the Danish system – "teaching for the test" would lead to more focus on test results than on the *National Common Objectives* of the teaching in the act of the *Folkeskole*. That is, too much focus on the tested profile areas and too little focus on creative, innovative and oral skills (which plays a significant role in the curriculum of the *Folkeskole*).

Who are allowed to know the results?

The main purpose of the testing system is to provide the teachers with a pedagogical tool – a tool which can help the teacher to analyse the proficiency level of the pupils and the level of the class. In order to reduce the incentive to "teaching to the test" and as precautionary measures against ranking of teachers, schools or local communities, it is forbidden by law to publish the items and the test results. Any test result obtained by a pupil, an average by a group of pupils, classes, schools, municipalities etc. are strictly confidential.

Only those, who for professional reasons need information about the results, are allowed to see them. All the results will be kept in a secured database. The database contains all the items used for testing the pupils as well as the answers that the pupils gave. Schools and municipalities are allowed to see and compare the results of the tests on different levels, according to their area of responsibility.

The teacher has access to detailed reports with information about the individual pupils' result as well as test results on class level. The headmaster is allowed to see the pupil's overall results, the class results and the results for the school. The local government/municipalities have access to result of the individual schools and this information aggregated to local government level.

The parents have to be informed by the school in writing about the results. For this purpose, the computer generates a verbalized report of the results for each pupil.

There is a strong tradition for parent involvement in the *Folkeskole*, and the test results should be used in the school's cooperation with the pupils and the parents in order to support each pupil in the best way possible.

What kind of response is given to the pupil, the parents, the teacher, the headmaster and the municipality?

As soon as possible after the test, the teacher will talk to the pupil about the result, and together they will plan the best way for the pupil to improve in the future.

The parents will be given the results for the different profile areas followed by a short explanation of the test results.

- The teacher is able to see the results for the whole class. The teacher will also be able to see which items a given pupil has answered as well as the result for the pupil.
- The headmaster is responsible for the teaching in his/her school. The headmaster is allowed to see the results for his own school and for the classes. The headmaster can also see the overall result for a given pupil. The headmaster will inform the school board about the results for the whole school.
- The municipality (which have the overall responsibility of the running and performance of the local schools) is allowed to see the results for the schools in the municipality. The local government/ municipality has access to the results of the individual schools and this information aggregated to local government level.

How can the teacher and headmaster use the results?

The results from the tests will enable the teacher to assess the proficiency of the pupils and of the whole class. The teacher can also see the difference in performance between the different profile areas.

The results will help the teacher to get at better overview over the pupils' proficiencies, in order to improve the teaching for the whole class and for the pupils individually.

The headmaster has the overall pedagogical responsibility at the school and therefore an obligation to guide and to coach the individual teachers in pedagogical matters. The test result should therefore be seen as a tool to pedagogic leadership.

The national results as an average score and a national profile of performance

The results from the first full-scale compulsory it-based tests will be used to define a scale for the schools - a reference for a national profile of performance. The purpose is to be able to monitor the overall development compared to the first year. Furthermore, the schools and the municipalities will be able to compare their results with the average results from the whole country.

When all the results from a compulsory test are registered, the mean performance of the pupils from the whole country and the distribution around the mean will be calculated. The distribution of the results will be separated by percentiles into 5 levels:

- 5 is given for the 10 % of the pupils that have the best results
- 4 is given for the next 25% of the pupils
- 3 is given for the 30 % of the pupils that are just around the mean
- 2 is given for the next 25% of the pupils
- 1 is given for the 10 % of the pupils with the lowest score

This distribution is calculated for each profile area and for the test as a whole. The results in the profile areas are called the National Profile of Performance.

The national results from the compulsory tests will be published annually. The results from the first test in a given subject will be used as a reference for the results in the following years. This will enable us to compare the results from the following years with the initial test and to see if the pupils' proficiency level is improving.

Correcting for differences in social background

The schools and the municipalities will be able to compare their results with the national results, – i.e. both the national profile of performance and the mean for the whole country. However, the background of the pupils in different schools is very different when it comes to socio-economic factors, which usually relate statistically to the pupils' test results. This will be taken into account and a statistical correction will be made. This correction will take into consideration factors as gender, ethnic background, parent's education and socio-economic status, etc. The corrected results are confidential, but will be given to the school and the municipality. This will make it possible to take the socio-economic factors into account when comparing the local results and the results from the whole country.

Plans for the future

After the completion of the pilot phase in 2009 and implementing of the improvements that are needed, the National testing system is planned to be launched in full scale in the spring 2010. It will provide information about the pupils' proficiency and knowledge to the Danish teachers.

But the system will not really show its worth until the schools have access to data for a couple of years. Then it will give the teacher an opportunity to monitor the individual pupil's progress. And it will give the teacher a possibility of advanced analysis of the progress of the class in a couple of ways that are not possible in any other system today anywhere in the world.

Furthermore, it will provide information to headmasters about the classes and schools that will enable new possibilities for pedagogical coaching and leadership.

It is though important to underline that this system – as any other test system - provides information about the pupil's proficiency, knowledge, attainment – not solutions on pedagogical problems. Even though in some cases the right interpretation of test results can deliver information that will tell the teacher, what would probably be the right thing to do with the individual pupil or the class. This kind of intelligent interpretation of test results could be the next area of development.

The author:

Jakob Wandall
Chief adviser
Skolestyrelsen (The Agency for the Evaluation and Quality development of Primary and lower Secondary education)
Snaregade 10 A
DK-1205 Copenhagen
Denmark

E-Mail: Jakob.Wandall@skolestyrelsen.dk

Jakob Wandall has been working with different parts of the educational system (adult- youth- primary and lower secondary education) within the Ministry of Education since 1994. Prior to that, Jakob Wandall has been working with research on social science, mainly educational issues. Since 2003 Jakob Wandall has been working with evaluation and assessment of student achievement in the primary and lower secondary education. One of the central projects has been the development and introduction of CAT (Computer adaptive testing) in the Danish school system.

Introducing Large-scale Computerised Assessment
Lessons Learned and Future Challenges

Eli Moe
University of Bergen, Norway

Abstract
The ability to use modern information technology is one of the aims stated in the Norwegian national curriculum (Utdanningsdirektoratet 2006). Statistics Norway (www.ssb.no) reports that the number of computers in Norwegian schools is increasing rapidly. The National Tests of English Reading for the 5th and 8th grades, commissioned by the Norwegian Parliament, are computerised. Developing computerised tests is a challenging enterprise since many decisions must be made. Once the tests are in place, and students know the test format, schools, teachers and pupils seem to like this form of testing. For test developers piloting then becomes easier, both with respect to the administration of the piloting and also because data is automatically generated. In this context, a number of interesting questions arise: Will the fact that the testing process has become somewhat easier, lead to more testing in schools? And if so – what might the consequences be? This article reports decisions made and lessons learned in connection with the introduction of large-scale computerised testing in Norwegian schools. In addition, based on the traditional Norwegian assessment culture and facts reported by some international educational surveys. I will consider long-term consequences of this form of testing.

Many researchers have reported that the testing and assessment culture in a country has an impact on society and teaching (Alderson and Wall 1993, 1996; Bailey 1996; Shohamy 2001; Wall and Horak 2006, 2007, 2008). In her article "The impact of society on testing" Cecilie Carlsen states that "there is a two-way relationship between testing and society: not only does language tests affect society; language tests are also affected by society" (Carlsen 2008).

Traditionally, the principles of democracy and equality have been an ideal for Norwegian society in respect to economic and cultural equality as well as educational equality. Consequently, there has been little testing as well as late differentiation between pupils in schools. The aim has been to give everyone an equal opportunity to perform well as well as allocating extra resources to pupils when necessary. Several studies have shown that Norwegian school children "like school a lot". Norwegians tended to think that Norway had the best educational system in the world. However, the PISA reading results of 2000 radically changed this view. Here Norwegian pupils performed averagely, a little behind Sweden and far behind the PISA winner Finland.

In the autumn of 2002 the newly elected conservative government decided to introduce national tests in Norwegian reading and writing, mathematics and English for the 4th, 7th, 10th and 11th grades. Pupils in the 4th grade are 9 to 10 years old, while pupils in the 11th grade are 16 to 17 years old. The original mandate asked the test developers to consider computerised testing. The development of these tests started in January 2003, and the University of Bergen has been responsible for the national tests of English.

The development process

Paper based tests or computerised tests?
The development of new computer technology, a proportionally high number of computers in the Norwegian population, high ambitions and a hope to improve educational standards led the Ministry of Education and Research commission to consider computerized tests for Norwegian school children in all skills – if possible.

From 1998 to 2003 the University of Bergen was responsible for developing the Norwegian version of Dialang (2008), the Internet-based language test in 14 European languages. This meant that the team developing the tests in English had some prior experience with computerised testing. But the fact that children were going to be tested, made this an even bigger challenge, since testing of children is, and has been, a controversial issue in Norway. We knew we wanted to test both receptive and productive skills, and decided to develop both reading and writing tests.

There are a number of studies comparing tests of L2 reading (reading in a second language) on computer and paper, however, few of these focus on the testing of children (Bennet 2003). Some of those which do, were conducted during the 1980s (Reinking & Schreiner 1985, Reinking 1988, Feldmann & Fish 1988). Increasing computer familiarity and an increasing number of computers call for further research into this issue. Johnson and Green (2004) study the impact of mode on student performance in mathematics, and find "no statistically significant difference in the overall difficulty of each test (computerized and paper)" (p.5), but suggest that "mode of assessment may influence the way that some children may think when answering questions" (p.9). After reviewing a large number of studies of comparability between computerized tests and paper & pencil tests, Sawaki (2001) concludes that "*the wide range of characteristics of participants, test tasks, test administration conditions, computer requirements, and the degree of control over extraneous variables observed in the studies reviewed in this article, as well as the scarcity of mode of L2 presentation research, make it difficult to draw conclusions based on these studies and to generalize the results to L2 reading assessment*" (Sawaki, 2001, p.51).

In the end we decided to develop computerised reading tests as well as traditional paper-and-pencil writing tests for English. In the following I will focus only the computerised reading tests. As this is something new in Norway, and few have much experience at all with computerised testing, the other teams developing national tests (mathematics, Norwegian reading and writing) have been reluctant to embark on the same journey. So far, the English tests have been administered four times on a full scale. From 2009 there will also be computerised national tests in mathematics.

Linear or adaptive tests?

The next step was to decide whether to develop linear or adaptive tests. While a linear test would present all pupils with the same items, an adaptive test would adjust the items to the each pupil's level of competence, i.e. each pupil would be presented with different tests. Since the tests were going to have both a pedagogical function as well as a reporting function, it was tempting to go for adaptive tests. An additional argument was that it would not be possible to administer the tests to all pupils in a grade (around 60 000 pupils) the same day due to server capacity and the number of computers in schools. The testing would have to take place within a time span of around two weeks. Adaptive tests would ensure that pupils, to some extent, would respond to different subsets of items, something which would prevent the content of the tests to become public knowledge very fast.

Making tests which were fully adaptive, seemed an overwhelming task, as we would need a huge item bank. We also felt that we to some extent lost control of the content of the tests if we voted for adaptive tests. We knew that an appropriate algorithm could ensure that all important facets of the test construct were included in the tests presented to the pupils. Still, we felt we needed a firm grip on the content of the test in order to convince teachers and parents that the tests measured what we wanted them to measure. Therefore, we asked ourselves whether it was possible to go for an in-between-solution, to have the best of the two worlds. Was it possible to have control over the content of the tests and at the same time have tests which more or less were adjusted to the pupils' level of competence?

In the end we decided to develop tests which were partially adaptive. Figure 1 shows the test format for the 11th grade. The levels A2 to C1 refers to the levels of the Common European Framework of Reference (2001).

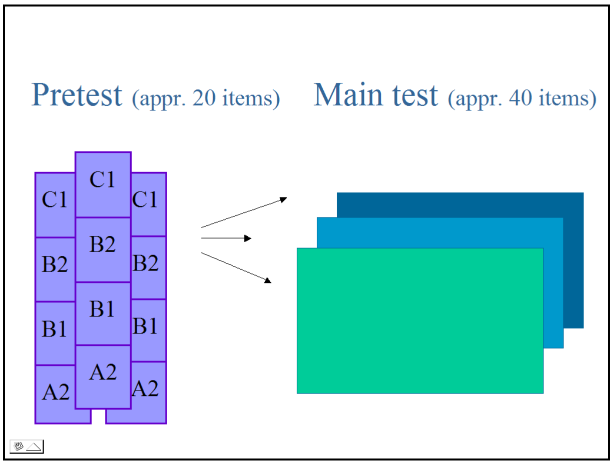


Figure 1: Test format - first edition

When a pupil logged in for a testing session, s/he was given one of three pretests randomly chosen by the computer. These three pretests had the same level of difficulty. Depending on the result of the pretest, the pupil continued to one of three main tests with different levels of difficulty. Pupils with a good command of English were given the most difficult main test, while weaker pupils are given the easiest main test.

Reactions – first round of testing

The first national tests were administered in the spring of 2004 to the 7th and the 10th grades. The tests in general, both paper tests and computer tests, received a strong negative public reaction amongst pupils in secondary school, as well as parents and teachers. Many secondary school pupils stayed at home the days the tests were administered, and their boycott was supported by parents' action groups. In their reactions they did not comment on the quality of the tests, but the negative reactions were brought about by egalitarian ideals: they were afraid that the testing and publishing of results would lead to more private schools and thus to greater differences between rich and poor. Teachers feared an increased work load, and parents feared more stress in school etc.

The computerized tests in particular received different types of reactions. The technical standards and sample tests were been public knowledge six months prior to the first round of testing. The schools were informed about this, and the sample tests were available on the net. Most schools had upgraded their computers to the recommended browsers and screen resolution. A few schools were not ready for computer testing. These schools either did not check their computers prior to the testing, and got media attention when they had problems loading the items on the screen. In addition, some schools had a relatively slow connection to the Internet, which meant that it took longer to load each task/item.

All in all – even though some schools had not been ready for computerised testing, in most schools the pupils in relevant grades participated. More than 100 000 pupils in the 7th and the 10th grade took the computerised tests of English. The next year (2005) the computerised tests were administered to four grades, the 4th, 7th, 10th and 11th grade, all in all around 200 000 pupils.

Many pupils and teachers liked the computer tests. The tests for the youngest pupils included many pictures, and pupils were for instance asked to click on items in a picture, to colour items in a picture, to click and drag items into a picture in addition to answering more traditional multiple choice items. Several teachers gave feedback that some of the young boys who normally had problems sitting still for a whole lesson were deeply absorbed in answering items when put in front of a computer. Many teachers were happy with the automatic scoring which did not increase to their work load.

An external team was hired by the Norwegian Examination Board to check on the quality of the tests. This team checked only the reliability of the main tests for all grades. The team's conclusion was that the reliability of the main tests was too low. They required a reliability of .85, and none of the main tests met this standard. What they did not consider, was the reliability of the pretests, as these were used for assigning pupils to main tests. The pretests had a reliability of around .90. Another thing the external team did not check was the standard error of measurement which was low for all tests.

Changes – second round of testing

Due to public reactions, media attention, public reports on the quality of tests as well as the developing teams answers to the reports, the *Ministry of Knowledge* (the Ministry had been renamed when the socialists won the election in 2005) decided not to administer national testing in 2006. When "new national tests" appeared in 2007 several changes were implemented:

	2004-2005	**2007 →**
Function of the tests	Pedagogical and reporting	Reporting
Subjects being tested	Norwegian reading and writing, mathematics, English reading and writing	Norwegian reading, mathematics, English reading
Reporting of results	Specific for each test	1 correct answer = 1 point
Grades being tested	4th, 7th, 10th & 11th	5th & 8th
Time of testing	Spring	Autumn

Table 1: National testing in Norway – changes implemented from first to second round of testing

In the first round the national tests had a pedagogical and a reporting function. For the test constructors, the double function represented a challenge. Considerations concerning one of the functions often conflicted with considerations for the other. For instance: The reporting function of the tests would benefit from clean scores that could easily be used as a basis for calculations of average scores across schools. The pedagogical function, on the other hand, required detailed information about pupils and their individual strengths and weaknesses. One score alone would not give useful information as to what aspects of language they needed to focus more on. The function of the "new" tests was limited to a reporting function.

In addition, the Ministry wanted fewer national tests and fewer grades being tested. The writing tests (both for Norwegian and English) were taken out of the national test pool, and the tests should be administered to pupils in the 5th and 8th grade. Pupils in the 10th and 11th grades had protested most fiercely in the first round of testing, and in the new system this problem was eliminated.

In the first two years of testing, the different national tests had reported their results in different ways. While the result in English were reported in terms of levels of the Common European Framework, the result of Norwegian reading and mathematics were reported in other ways. The Ministry found this too complicated and therefore decided that all the national tests had to report their results in the same way: each correct answer would give one point. The points reported to the pupil would indicate how many items the pupil had answered correctly.

The adaptive tests were abolished, and the team was told to make linear tests. As a result of this we now develop three parallel linear tests, both for the 5th and the 8th grade.

When a student logs in to take the test, s/he is sent automatically to one of the three parallel tests. The tests have the same level of difficulty and discrimination indices, and are comparable in respect to empirical difficulty, number of items, item formats, themes etc. Table 2 and 3 are examples of how we, after piloting of items, document that the tests we have developed, are parallel. The example is taken from the 5th grade tests.

Item format	Version 1	Version 2	Version 3
Click item	3	3	3
Click and drag	4	5	4
Colour	3	3	3
Click picture	6	6	6
Click text	7	7	7
Gap filling	4	2	6
Multiple choice	2	2	3
Who could say	9	10	6

Table 2: National tests of English 5th grade – number of items and item format - all versions

	Version 1	Version 2	Version 3
Number of items	38	38	38
Reliability (α)	0.924	0.921	0.925
Mean raw score*	21.7	21.5	22.1
Standard deviation (raw score)*	8.4	8.3	8.7
Mean difficulty	57%	57%	58%
Mean discrimination	0.50	0.49	0.50

Table 3: National tests of English 5th grade–statistics (* Based on simulation data (n= 1000))

2008 – Status

By November 2008 the computerised tests of English had been administered on a full scale four times. The schools have upgraded their computers and Internet connections, and everything seems to be running very smoothly. A report from 2007 states that

- more than half of the students say they like the computerised tests of English "very well"
- more than half of the pupils say the tests are easy, while the results show that this is not true
- there are significantly more boys than girls that think the tests are easy
- when the pupils are asked whether they know the item formats well – 31% disagree (in contrast 27-28% of the pupils say the same about the paper based tests in Norwegian reading and mathematics)
- there are significantly more boys than girls that say they know the item formats well

- some pupils say there are many texts to read, and that is hard to read from a screen
- one of 3 say they are not able to show their ability of English through these tests, which is, of course, partly true since its only English reading being tested
- There are no significant differences between girls and boys regarding test results.

In general things have calmed down. The national tests do not get much attention from the media in connection with test administration. There is some attention when the results are published around a month after the administration. The results reveal some differences for instance between urban and rural areas. But now people have started asking "why" these differences occur, and this may turn out to be a constructive way of handling these issues.
From a test developer's point of view the development of large-scale computerised tests has been, and still is, a very interesting and challenging process. Educational administrators, people building the technical platform and test developers sometimes seem to live in different worlds, and it may be very challenging for persons from one of the groups to communicate with those in another group. To be successful it is very important to ensure that the different stakeholders are able to communicate with another. When developing these tests, it was necessary to work closely with groups as diverse as teachers, the staff building the technical platform, a statistician, persons with an e-learning background, an artist drawing pictures used in the tasks for the youngest pupils. Norwegian schools have been very cooperative and positive to taking part in piloting of items. School administrators and teachers say that taking part in piloting is the best way to check their computers and Internet connection, and that it is a clear advantage that the computerised tests don't add to the teachers' work load.

Future challenges

So far, the English tests have been testing reading alone. From 2009 we are adding vocabulary and grammar. Vocabulary and grammar will not be tested in isolation, but in a context. We also hope to include listening items in a few years time. The reason for doing this is to measure more facets of the pupils' ability of English.

And what about the productive skills? What about speaking and writing? I doubt pupils will say that they are able to show their ability of English if the skills of speaking and writing are not included in the tests. So far, our main focus has been to develop tests which are relatively easy to administer and to score. In the next round we should perhaps also assess whether it is possible to include productive skills.

Concluding remarks
The PISA 2000 reading results brought about major changes in the Norwegian educational system. After a short period of protest, people seem to have accepted the new testing system, and many say the tests make teachers and schools think more profoundly about assessment issues and how to improve learning. The PISA 2000 survey also had some positive findings which did not receive much attention in the media: Norwegian children obtained high scores on social well-being at school. This finding is supported by a UNICEF report from 2007: An overview of child well-being in rich countries. According to that report more than 40% of Norwegian school children (11, 13 and 15 years old) say they "like school a lot". Of the OECD countries Norway is at the top, and the PISA winner in respect to proficiency, Finland, is at the bottom. Only 7% of Finnish children say they "like school a lot". Clearly then, the anti-elitist, unitary school system has at least some positive outcomes.

Since the PISA 2000 reading results, there has been more focus on testing and documentation of learning outcomes. New diagnostic computerised tests of English are being developed for the 11th grade. Computerised testing makes testing easier to administer, and an increased number of tests are on their way. The big question is whether future Norwegian pupils will continue to report that "they like school a lot".

References

Alderson, J. C. and Wall, D. (1993). Does Washback Exist? In Applied Linguistics 1993 Vol. 14, No. 2: 115-129.

Alderson, J. C. and Wall, D. (1996). TOEFL preparation courses: a study of washback. In Language Testing 13/3: 280-297.

Bailey, K. (1996). Working for washback: a review of the washback concept in language testing. In Language Testing 13/3: 257-279.

Bennet, R.E. (2003). Online Assessment and the Comparability of Score Meaning. Paper presented to International Association for Educational Assessment annual conference, Manchester, October 2003.

Carlsen, C. (2008). The impact of society on testing. In Research Notes, Issue 34, November 2008. http://www.cambridgeesol.org/rs_notes/

Dialang. (2008). http://www.dialang.org/english/index.htm

Feldmann, S.C., & Fish, M.C. (1988). Reading comprehension of elementary, junior high and high school students on print vs. microcomputer-generated text. Journal of Educational Computing Research, 4, 159-166.

Johnson, M & Green, S. (2004). On-line assessment: the impact of mode on student performance. Paper presented at the British Educational Research Association Annual Conference, Manchester, September 2004.

Reinking, D. (1988). Computer-mediated text and comprehension differences: The role of reading time, reader preference, and estimation of learning. Reading Research Quarterly, 23, 484-500.

Reinking, D. & Schreiner, R. (1985). The effects of computer-mediated text on measures of reading comprehension and reading behaviour. Reading Research Quarterly, 20, 536-552.

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. Language Learning & Technology, vol. 5, Num. 2, 38-59.

Shohamy, A. 2001. The Power of tests. A Critical Perspective on the Uses of Language Tests. London, Longman.

Statistics Norway. (2008) www.ssb.no

UNICEF (2007). An overview of child well-being in rich countries. http://www.unicef-irc.org/publications/pdf/rc7_eng.pdf

Utdanningsdirektoratet (2006). Læreplanverket for Kunnskapsløftet. Oslo: Utdanningsdirektoratet. http://www.udir.no/templates/udir/TM_Tema.aspx?id=148

Kavli, H. (2008). Nasjonale prøver 2007 – Brukernes evaluering av gjennomføring. http://www.utdanningsdirektoratet.no/templates/udir/TM_artikkel.aspx?=3413#

Wall, D. and Horak, T. (2006). The TOEFL Impact Study: Phase 1. The Baseline Study. In TOEFL Monograph 34, NJ: Educational Testing Service.

Wall, D. and Horak. (2007). Using Baseline Studies in the Investigation of Test Impact. Assessment in Education 14/1: 99-116.

Wall, D. and Horak. (2008). The TOEFL Impact Study: Phase 2. Coping with Change. TOEFL iBTResearch Series, No. 05. Princeton, NJ: Educational Testing Service.

The author:

Eli Moe
University of Bergen
Soendre Steinkjellersmug 7
5003 Bergen
Norway

Telephone: + 47 5558 2216 (work)
+ 47 9119 6514 (cell)

E-Mail: Eli.Moe@lle.uib.no

Eli Moe is a former language teacher with a special interest in second language acquisition who for the last 12 years has worked with a team developing language tests at the University of Bergen / The University of Bergen's Research Foundation (Norway). She is working with a team developing language tests in Norwegian for adult immigrants (traditional paper & pencil tests), and she is leading the work of developing of national computer-based tests in English for Norwegian school children. Her research interests are within second language acquisition, standard setting and computerised testing.

Large-scale Computer-based Testing of Foreign Language Competences across Europe

Technical Requirements and Implementation

Jostein Ryssevik
ideas2evidence ltd, Norway

Abstract

This article presents the software requirements of the European Survey on Language Competences (ESLC) – a large-scale international survey initiated by the European Commission to study the foreign language skills of European school children. The emerging open source software platform developed by the SurveyLang consortium is also described. The ESLC will be conducted in all or most EU member states and is aimed at pupils of lower secondary education studying one or both of the two most taught foreign languages in their country. The sampled schools will have the opportunity to choose between computer-based administration or a paper-and-pencil based equivalent. It is, however, a goal of the project to maximize the number of schools that choose the CBT-version and the software platform is designed to support this goal. The article describes the demanding technical requirements deriving from this dual-mode approach and the fact that the study is based on a complex incomplete block design where the delivered tests will be matched to the proficiency levels of the students. The software support for the item writing process is also described in this article, as well as the planned solution to the strict security requirements that always go hand in hand with international assessment surveys like ESLC.

In order to develop a *European Indicator of Language Competence*, the European Commission has initiated a large-scale assessment survey to be conducted for the first time in the first quarter of 2011. The contract to develop, manage and analyse the output of the *European Survey on Language Competences*, was awarded to SurveyLang - a consortium of European partners headed by Cambridge ESOL and involving also among others the National Institute for Educational Measurement (CITO) and Gallup Europe. The development of the software platform is managed by Gallup Europe.

ESLC – basic parameters

According to the European Commission's plans, the survey should be conducted in 32 countries across Europe, although it is still not decided whether all the invited countries will take part in the first round. The ESLC will test the competencies of European pupils in five languages - English, French, German, Spanish and Italian - but only the two most frequently taught languages out of these five will be tested in each country. The first round of the survey will focus on testing three language skills: listening, reading and writing. Measurements of speaking skills might be added in future rounds.

The target group of the study are pupils in the last year of lower secondary education or the second year of upper secondary education who are studying one of the two most taught foreign languages in their country. Due to major differences across countries when it comes to the introduction age of foreign language teaching, especially regarding the second foreign language, it is anticipated that the target group will have to include upper secondary pupils in some countries. In each country 1500 students for each of the two chosen languages will be randomly sampled to take the test. Each student will only be tested in one language and only in two of the three skills mentioned above.

In order to increase the precision of the tests and to avoid fatigue or boredom effects, it has been decided to introduce an element of targeting. As full-blown adaptive testing is hardly feasible and probably not desirable in a survey like the ESLC, a hybrid design has been developed based on a short routing test taken prior to the main survey. The routing test will classify the pupils in three proficiency levels. The information from the routing test will subsequently be used to allocate students across tests at different difficulty levels in a linked design that makes sure that each pupil receives at test that is targeted to his or her proficiency level. Combined with the principles of an incomplete design where each pupil will only will receive a portion of the test material that matches his or her proficiency level, targeted testing involves an extremely complex logistical scenario when it comes to the assembly and dissemination of unique testing sequences for each single student.

In order to reduce the burden on the participating countries, it has been decided to develop the ESLC for computer-based testing. However, given the variations in technological preconditions and computer skills across schools and countries, a paper-and-pencil based equivalent will also be offered. Besides the methodological challenges, this dual-mode design also increases the logistical and technical challenges. As will be described below, it necessitates a software platform that can support the administration of paper-and-pencil-based tests as well as of the computer-based tests in a coordinated way. An explicit objective is also to develop the software platform in such a way that as many schools as possible will be in a position to choose the computer-based alternative.

Requirements

The technical and functional requirements of a software platform designed to support a large-scale survey operation like the ESLC are demanding. At a high level, the software platform should:

- support all the various stages and roles in the development and implementation of the survey (see Figure 1),
- enable the automation of error-prone and expensive manual processes,
- be flexible enough to handle the variety of test item types used by the survey,
- support the implementation of the complex incomplete block design described above,
- meet the high security requirements of international assessment surveys like the ESLC
- reduce the technical and administrative burden on the local administrators to a minimum
- run on existing hardware platforms in the schools
- be an open source platform available for free use by other actors (an explicit requirement in the terms of reference of this contract)

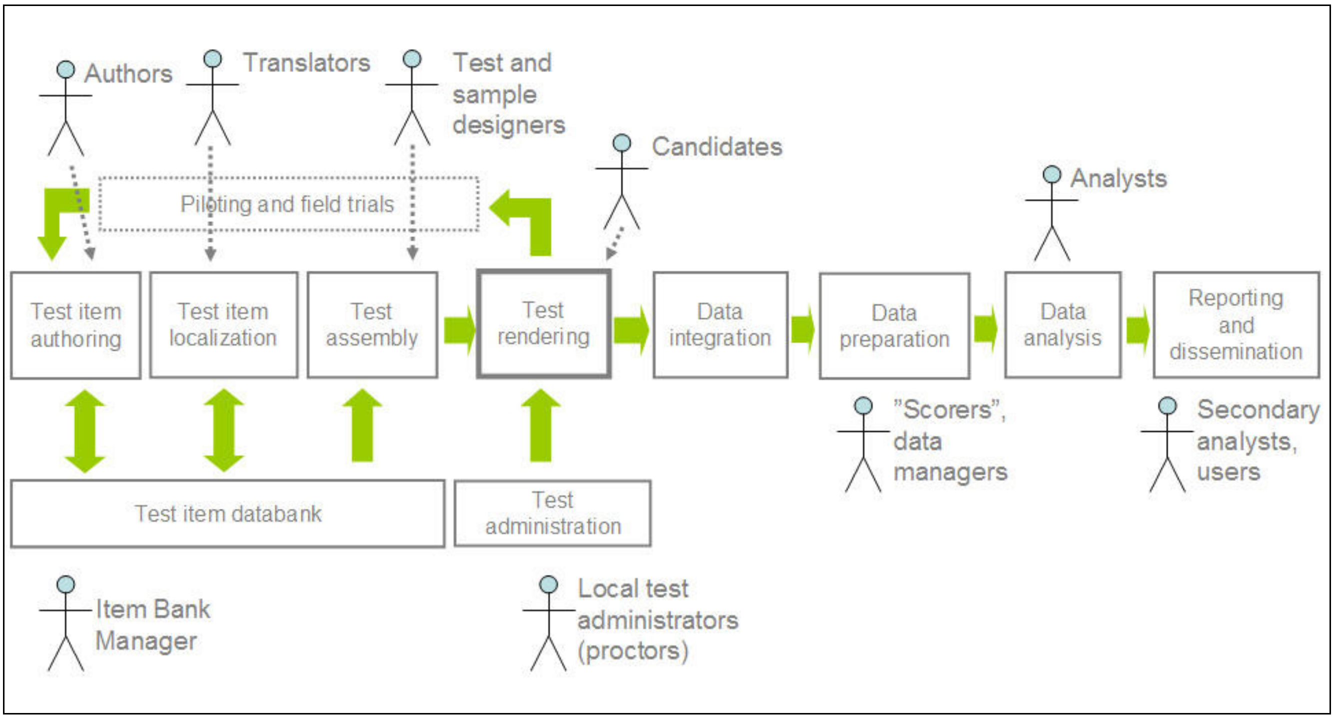


Figure 1: Stages and roles in the design and delivery of the survey

In terms of functionality, the following tools and components are needed:

- Test-item authoring, editing and preview functionality supporting a system of distributed authors scattered around Europe.
- Test-item databank functionality providing efficient storage, management and version control of test-items. This tool should also encourage visibility and sharing of recourses between the various roles associated with the stages of the test-item life-cycle.
- Test-item translation functionality, supporting the localization of test-items, instructions and accompanying questionnaires to national languages.
- Test construction functionality, supporting the assembly of individual test-items into complete test sessions (compatible with the targeting and block substitution structure required by the overall design) as well as the allocation of students across tests at different levels.
- Test administration functionality supporting the management of respondents and test-sessions at the school level.
- Test rendering functionality supporting efficient and user-friendly presentation of tests-items to respondents as well as the capturing of their responses
- Data integration functionality supporting efficient assembly of response data coming from the participating schools.
- Data preparation functionality supporting all tasks related to the preparation of data files ready for analysis, including support for manual marking/scoring of open ended items.
- Data reporting functionality supporting online access to analytical results as well as download of files for statistical analysis.

Architecture

The high level architecture of the software platform that has been designed to provide this functionality can be seen in Figure 2. The platform consists of a central Test-item databank interfacing three different tools over the Internet: 1) a Test-item authoring and editing tool, 2) a Translation management tool, and 3) a Test assembly tool. As a whole, these three distributed tools, plus the Test-item databank, are designed to support the central test development team in their efforts to develop and distribute the language tests.

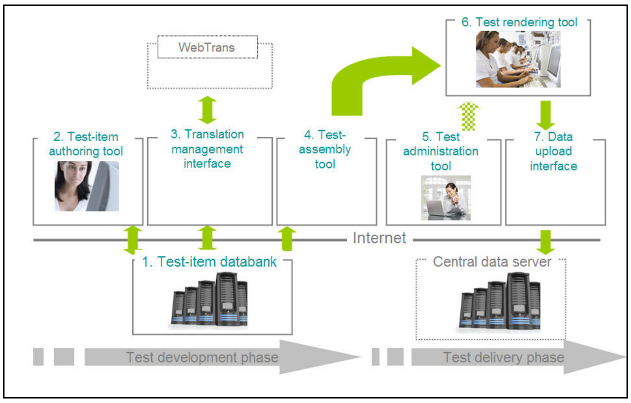


Figure 2: High level architecture

To support the test-delivery phase of the project, another set of tools will be provided. These are 1) a Test-rendering tool to be installed on the test computers in all the schools taking CB-testing and 2) a Test-administration tool supporting the various tasks of the local test administrators. Test rendering will take place on computers which are disconnected from the Internet. The Test-administration tool will however, provide an interface that will allow local test administrators to upload the collected data to a central database over the internet.

In the following paragraphs we will describe some of these tools in further detail.

Test-item authoring

The test-items of the ESLC survey will be developed by an expert team of 50+ test-item authors distributed across Europe doing their work according to specifications and guidance provided by the central project team. Items will move through various stages of a predefined life-cycle including authoring, editing, vetting, adding of graphics and audio, pilot-testing, field-trial etc., each stage involving different tasks, roles and responsibilities.

The Test-item authoring tool is designed to support this distributed and fragmented development model. It is also designed to allow non-technical personnel to create tasks in an intuitive way by means of predefined templates for the various task-types that will be used in the survey. At any stage in the development, a task can be previewed and tested to allow the author to see how it will behave and look when rendered in a final test. The authoring tool will also support the capture and input of all the metadata elements associated with a task, including comments and descriptions, versioning metadata, test statistics etc.

The tool is implemented as a rich client by means of technologies like Adobe Flex and Adobe Air. This provides a very user-friendly and aesthetically pleasing environment for the various groups involved in the development of the tasks.

Test-item databank

The Test-item databank is the hub of the central system providing long-term storage, version control and management of test-items and their associated metadata and rich media resources. Test-items will be uploaded to the databank by the test-authors to be seen and shared by others. When, as an example, a task has reached a stage in the development where an audio file should be added, the person responsible for this stage will download the task, read the audio transcript, create and attach the soundtrack and load the task back up to the databank. The databank will include a version control mechanisms keeping track of where the task is in the lifecycle as well as a secure role-based authentication system, making sure that only authorized personnel can see or change a task at the various stages in the life-cycle.
The Test-item databank is implemented in Java on top of Apache Tomcat and MySQL communicating with the various remote clients through Adobe Blaze DS.

Translation management

It goes without saying that a software platform developed for foreign language testing will need to be genuinely multilingual. Not only will equivalent language tests be developed in the five target languages. User guides, menus, navigation elements and questionnaires will in addition be offered in all the national languages of the countries where the tests are taken. Each concrete test presented to a respondent will thus have two different languages; a target language and the national language of the location where the test takes place. This requires efficient language versioning and text string substitution support. It also requires an efficient, robust and scientifically sound translation management system.

Gallup Europe – one of the main partners of the Surveylang consortium - has already developed a translation management system called WebTrans for their large-scale international survey operations, amongst other the Commissions' Flash Eurobarometer project. This WebTrans system supports central management of translators scattered all over Europe and a model of forward and back translation similar to the one that will be used for ESLC. The consortium has decided to make use of WebTrans and to create an interface between that system and the Test-Item Authoring tool.

Test assembly

The Test assembly tool is without doubt the most sophisticated piece of software in the Surveylang platform. The tool is designed to support three important functions (see Fig. 3):

1. the assembly of individual test items into a high number of complete test sequences,
2. the allocation of students across these test sequences according to the principles and parameters of the predefined survey design (see above),
3. the production of the digital input to the computer-based Test rendering tool, and
4. the production of the digital documents that will be used to print the paper-based test booklets.

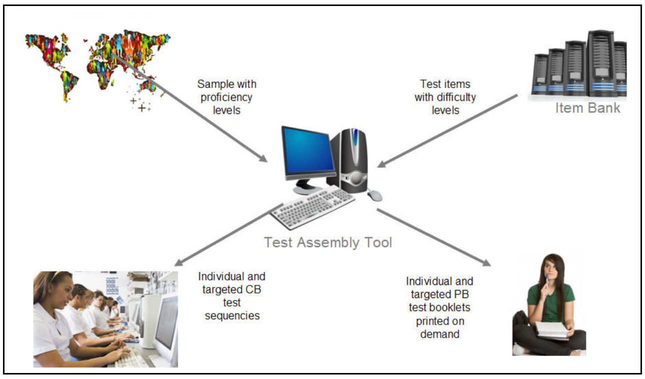

Sample with proficiency levels
Test items with difficulty levels
Item Bank
Test Assembly Tool
Individual and targeted CB test sequencies
Individual and targeted PB test booklets printed on demand

Figure 3: The roles of the test assembly tool

The assembly of test items into complete test sessions takes place in two steps. The first step is handled by the test designers and involves the construction of a high number of blocks (of approximately 15 minutes testing time). Each of these blocs contains tasks focusing on a single skill and is assigned one of four difficulty levels. The second step is fully automated and involves the assembly of blocks into test sequences, each containing two skill sections (of approximately 30 minutes testing time) plus a student questionnaire that will be administered to all the students. The difficulty level of the test sequences will directly derive from that of the blocks which they are composed of. The test sequences are assembled according to the basic rules of the survey design which will be formalized in XML and interpreted by the Test Assembly tool. The tool will subsequently be able to handle changes to the design without any reprogramming.

In order to allocate students across test sequences, the Test assembly tool will also need to be able to make sense of detailed information about each single sampled student; especially the student's proficiency level (from the routing test), the target and national language of the student, the student's school and whether this school is taking a computer-based or a paper-based test. Based on this information and following the principles of targeted testing, the Assembly tool will randomly allocate students across the group of available test sequences matching their respective proficiency level.

The last task of the Test assembly tool is to produce the individualized test material for each single student, both for computer-based and paper-based testing. In the former case the material will be produced as a package of tests in a predefined XML-format ready for use by the Test rendering tool. In the latter case, the Test Assembly tool will produce a package of digital documents (PDF), each document containing the paper-based form of a named student. In both cases, digital packages will be produced for each single school in the sample based on the combination of target and national language relevant for that school.

What makes this an innovative approach is the fact that both modes (computer and paper) are served from the same source and the same system. This is reducing manual work and hopefully also manually-induced errors. According to the experiences of other international surveys involving complex test construction, printing of forms and the process of handing the right form to each individual student, are the most complicated and error-prone processes in the administration of the survey. By offering a solution where the package of individualized and named forms can be printed on demand for each single school, we hope to reduce these problems to a minimum.

Test rendering

One of the critical challenges related to computer-based delivery of assessment tests is, in general, security. On the one hand, it is crucial that the content of the tests are protected from disclosure before and during the testing period. On the other hand, it is of utmost importance to create testing environments that are as equal as possible for everyone and where the students are protected from external influences of any sort (like access to the web, chatting channels, digital dictionaries etc.) while taking the tests. For the latter reason the test will have to take place on test computers that are disconnected from the net and where the desktop of available tools and software is fully controlled by the test administrators.

If the tests could have been taken on dedicated test computers brought into the schools for that very purpose, the problems would have been trivial. However, in a scenario where all tests will be taken on the schools' existing hardware platforms, this is more of a challenge. The solution that we are opting for is to boot the test computers from USB memory-sticks or CDs, including a minimum-size operating system (a Linux variety), the test rendering tool and the complete package of tests. In this way we will be in full control of the local desktop, we can block the access to Internet and we can monitor that the tests actually are taken in this closed and controlled environment. We are aware of many technical obstacles that will have to be overcome for this to happen smoothly in every technical environment. We will however use pilot testing and the upcoming field-trial to map and develop workarounds for these obstacles.

The test rendering tool will be implemented in Adobe Flex and run in the Adobe AIR runtime environment. It is designed to support the rich multimedia test format generated from the Test assembly tool.

The administration of the tests at school level will be managed through the Test administration tool. This tool is also including the functionality to upload the captured data to the central data server. In order to reduce the requirements for the local hardware platform to a minimum, this part of the Surveylang software platform will not be dependent on a local area network (LAN). It is expected that this approach will increase the number of schools that are in a position to take the computer-based version of the tests. As a side effect, we also expect a lower number of failures due to local technical problems.

Open Source

The SurveyLang software platform will be developed as open source. This means that the platform will be free and open for use and extension by others. As soon as the development has reached a certain level of stability, the code and documentation will be made public and distributed under a standard open source license. The project is also based on open source development strategies, including a will to mobilize external developers and testers and to collaborate with other projects. As much as possible the development is based on existing open source frameworks, tools and components. We are also following open standards and using technologies that support an open source development model.

Conclusions

The software platform currently being built by the Surveylang consortium is designed from the bottom up to meet the complex requirements deriving from large-scale international assessment surveys. This includes support for the standard questionnaires that are normally part of these surveys. The software is built to support language testing, but could easily be extended to support other task types and subject domains. It is built around a generic task structure and life-cycle model that maps to international standards like QTI and DDI.

The implementation of the software started in September 2008 so there is still a way to go before the first versions of the software will be released. A complete suite including all the tools described in this article will be ready for the ESLC field trial in the first quarter of 2010.

The author:

Jostein Ryssevik
E-Mail: Jostein.Ryssevik@ideas2evidence.com
Phone: (+47) 91817197
Web: www.ideas2evidence.com

Jostein Ryssevik has twenty-five years of experience in social science data management. He has headed a long range of international software development projects within the fields of statistics, data analysis and knowledge management. From 2001 to 2005 he served as technical director of the UK-based software company Nesstar Limited developing tools for Web-based publishing and analysis of statistical data. Ryssevik is an expert on statistical metadata. He has played an active role in the development of the DDI metadata standard and is one the founders and managers of the Open Data Foundation, a non-profit organization dedicated to the adoption of global metadata standards and the development of open-source solutions promoting the use of statistical data. He is currently managing the independent consultancy company ideas2evidence and is responsible for the development of the software platform for the European Survey on Language Competences.

Delivery Platforms for National and International Computer-based Surveys
History, issues and current status

Sam Haldane
Australian Council for Educational Research (ACER), Australia

Abstract:
This paper traces the history of systems developed and used in a selection of large-scale computer-based surveys. It addresses the issues raised at each of the development stages and the various solutions proposed and subsequently implemented. Those issues include security of software and data, specifications of the hardware and software used, perceptions of network administrators and test administrators, economics of delivery and data capture, collation and marking, and creating and presenting material in different languages. The methods and delivery system used in PISA 2006 are critiqued together with those trialled for PISA 2009. In the latter case solutions to issues raised in the field trial will be addressed. Finally, the current status of delivery systems for large-scale international and national surveys will be measured against a perceived ideal for achieving the potential promised by computer-based assessment.

Computer-based assessment is becoming more and more common. As technology improves, the requirements of computer-based assessment delivery systems are expanding and becoming more demanding. The Australian Council for Educational Research (ACER) has been involved in several large-scale computer-based surveys in the recent years. Each survey had different objectives and requirements, and because of this different systems were developed and used in the field.

PISA Computer-based Assessment of Science (CBAS)

The PISA Computer-based Assessment of Science (CBAS) project in PISA 2006 was the first time a computer-based component was included in the PISA project. The main aim of CBAS was to create and administer a test that assessed students' science literacy using a computer-based environment. CBAS leveraged the computer-based delivery method to add value that could not be achieved using the traditional paper based test. To achieve this, rich elements like video, audio and interactive simulations were to be used, reducing the overall reading load of the test.

Objectives

Comparability of the test between students and participating countries was a major objective of PISA CBAS. The paper based PISA test has very well defined and strict standards with regards to the translation and presentation of items, to ensure that students in different countries have a very similar experience when taking the test. Rigorous translation and verification procedures, item review, and standards for print quality are just some of the steps taken to ensure this comparability in the paper-based test. This high standard of comparability was to be taken over to CBAS to ensure that students taking the test in countries and schools with better computer equipment did not have a better experience and hence find the test easier than students taking the test in countries and schools that were not so well equipped.

Reliability was another major objective for CBAS. Failure of a test session is quite costly, both in terms of test administrator time and loss of data, so the system developed should be as reliable as possible, building on high quality and tested components. In the case of a test failure, the system should have data recovery mechanisms to preserve whatever data was collected in the session before the failure.

The CBAS system needed to support fully translatable software and items, due to the international nature of PISA. All elements of the software needed to be translatable, as well as all text elements within the items, including dynamic text contained in items with interactive simulations. Right-to-left scripts such as Hebrew and Arabic also needed to be supported, meaning that the software itself needed to support mirroring of the interface; with user interface components that would be on the left for the English version should be on the right for the Arabic / Hebrew version.

The main way that CBAS added value to the test was by utilising media such as video, audio and interactive simulations. The system requirements were conceptualised with this in mind, and the fundamental technology used to build the system was chosen with this and the afore-mentioned objectives in mind.

Requirements

The fundamental requirements of the system developed for PISA CBAS reflected the main objectives mentioned above. Security was also a concern. As all PISA items are secure, it was a requirement that no item material was left on student or school computers after the test sessions. Students should also not be able to compromise a test session by terminating the CBAS delivery software.

The hardware and software used was required to be affordable at the time of the field trial (which was in 2005), but still able to facilitate the rich content that was a main objective of the project. Where possible, free and open source software should be used to avoid licensing costs.

The system was required to be as easy as possible for test administrators to set up and use. Generally speaking in most countries, PISA test administrators are retired or semi-retired teachers with limited technical knowledge. While the Consortium can recommend that test administrators with substantially more technical knowledge be used for PISA CBAS, the reality is that the same test administrators used for the paper-based test would be used for CBAS. Therefore all effort was to be made to design the system to be as user friendly as possible.

Implementation

With the main objectives and requirements in mind, the CBAS system was implemented to work as a client-server model. One computer was used by the test administrator to control the test session (the server), and this computer was networked to five other computers that students used to take the test (the clients). The response data from each student was transmitted over a local area network back to the test administrator's computer (the server). This model was chosen to make the data collection procedures easier, and to allow the test administrator to centrally control the test session from one computer, making the session administration easier.

To ensure the highest comparability possible, it was recommended that participating countries use mini-labs of six computers that the test administrator set up in advance, and carried in to the schools. This increased the logistical requirements of the study but minimised the set up time per school, and ensured greater comparability. Several specific laptop models were recommended, all of which had the same technical specifications such as screen size and resolution, CPU speed and memory capacity.

The Java programming language was chosen due to its widespread support, and abundance of open source libraries available. The Java Media Framework (JMF) library was used to provide support for video and audio. Java has good support for internationalization and localization, which was a major requirement of CBAS due to the international nature of PISA. Java also has very good libraries for user interface design, and networking. To implement items with interactive simulations, Adobe Flash was chosen. A proprietary third party library was used to integrate Flash content into the Java based delivery applications.

Microsoft Windows was chosen as the supported operating system because the majority of new laptops at the time came with a version of Windows pre-installed. The delivery system was implemented as two desktop applications that were installed onto the CBAS laptops. One application was for the test administrator (the server application) and the other was for the students to take the test (the client application). All item content such as video, audio and Flash was installed on the client computers along with the client application to reduce the network bandwidth required.

Initially, wireless networks were recommended in order to ease the technical setup required to be done by the test administrator. This proved to be problematic in many countries as interference from other wireless networks, microwaves and even airport radar systems caused some session failures in the field trail. Therefore, for the main study wired networks were recommended as a more reliable technology.

Economics
A major concern of many participating countries was the large cost of implementing CBAS. Only three countries participated in the main survey of CBAS. The main source of this large cost was the purchase or hire of mini-labs of six computers and networking hardware. Logistics were also a major concern, especially in cities like Tokyo, Japan where it was unfeasible for test administrators to travel by car, and they were also unable to carry the required equipment on the subway. The total control over the delivery system hardware and software did result in a very low test session failure rate, however.

Data capture was largely automated and semi-centralised on the test administrator's computer. Each test administrator was required to copy the data onto a device (e.g. CD or USB stick), which was then sent to the national centre where the national data was consolidated.

Marking of the CBAS items did not cost anything, as due to the nature of the CBAS items, marking was totally automated. No open response items or any other items requiring human marking were administered in PISA CBAS.

Translation of the CBAS items was quite cost effective, mainly due to the low word count of the CBAS items themselves. Custom software was created to facilitate the ease of translation of the CBAS items. The translation software was required to be installed on translators' computers, but after the initial installation step translation was relatively easy. An online translation website was developed to facilitate the download and upload of translations, and also to support the CBAS translation procedures.

PISA Electronic Reading Assessment (ERA)

The PISA Electronic Reading Assessment (ERA) in PISA 2009 is the second time a computer-based assessment component has been included in a PISA cycle. In 2009 the focus is on electronic texts as opposed to science (as in CBAS). A typical item contains a simulated website environment along with a question relating to the website content.

Objectives
Low implementation cost was the main objective that shaped the implementation of the ERA systems. The high cost of implementing CBAS meant that many countries were unable or unwilling to participate, so in order to have more countries participating in ERA, a solution that uses existing school information technology infrastructure was sought. A compromise of comparability is a result of using existing school infrastructure, as different schools around the world have different computer systems, with varying screen sizes and resolutions, CPU speeds and RAM capacities.
Logistics were a major concern with many countries for CBAS. A solution that didn't require test administrators to carry around laptops was a main objective of ERA.

The nature of the *Electronic Reading Assessment* meant that a complex hypertext environment was needed. A complex amount of interactivity within the websites was a main objective of the project, with multiple websites existing in the one unit, and interactive features like email and blogs.

Requirements
The requirements of ERA were quite strict in that not much existing infrastructure could be assumed. Because ERA was required to use existing school infrastructure, things such as inter-/intranet connectivity and host operating system version could not be relied upon. Due to differing security policies in schools around the world, it could not be assumed that software could be installed on the school computers.

Computers in schools around the world vary greatly in specifications; therefore the ERA system was required to run on the lowest common denominator hardware, assuming a reasonable set of minimum system requirements.

Implementation
The ERA system was implemented as a bootable CD (with a bootable USB option available in the main survey, commencing early 2009). A USB memory device was used for data collection. The bootable CD contained all of the software required to deliver the ERA test, which meant that no software was required to be installed on the school computers at all.

The bootable CD / USB contains a Linux distribution which has been customised for ERA, with the appropriate software, fonts and Input Method Editors (IMEs). The system uses a standard browser and web server, which run directly from the CD. The TAO framework (http://www.tao.lu/) was used as the base technology to deliver the test, with some custom Flash components used to deliver the simulated hypertext environment. TAO is a joint initiative of CRP Henri Tudor and the University of Luxembourg which provides a general and open architecture for computer-based test development and delivery.

The system was designed to use the minimum possible amount of hardware resources, so that it could run on the maximum amount of existing school computers possible. Care was taken to optimise CPU and memory usage where possible.

Most computers are configured to boot from a bootable CD or USB memory device when there is one present. Some computers require a small procedure to be undertaken in order to enable this functionality, however. This procedure involves changing the boot sequence inside the Basic Input / Output System (BIOS) of the computer. This is a somewhat technical procedure, but is required in order for the ERA system to run.

Economics

The design of the ERA system to use existing school information technology infrastructure ensured a low cost of delivery relative to CBAS. However, a relatively high amount of session failures occurred. This was due to some existing school computers not meeting the minimum hardware requirements of the ERA software, or the school hardware not being compatible with the Linux distribution used.

Data capture was higher cost with ERA, as each computer used collected the student response data on a USB memory device. The test administrators then were required to consolidate the collected data by copying the captured data files from each USB device onto a computer (often a laptop carried to each test centre).

Expert (non-automated) marking was required for some ERA items. Custom online marking software was developed to facilitate distributed marking, which also centralised the marking data collection.

Translation for ERA was quite costly due to the high word count of the items. A custom translation management system was developed to facilitate the download / upload of translations. XML Localisation Interchange File Format (XLIFF), a standard supported by many open source and commercial translation software packages was used in order to reduce costs. This enabled translators to use software that they are used to, and also eliminated the need to develop a custom translation application (as was done for CBAS).

National Assessment Project – Information and Communication Technology Literacy Assessment

The Information and Communication Technology Literacy project (ICTL) was commissioned by the Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) in Australia. The study was an Australia wide assessment of ICT literacy in year six and year ten (twelve and sixteen year old) students. It aimed to measure students' abilities 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.

The ICTL items themselves are very rich. A lot of items contain emulated application environments such as word processors, presentation preparation applications and photo organising applications. The test itself was delivered using remote desktop technology.

Delivery Model

The ICTL project utilised three delivery models, depending on what existing infrastructure each school included in the study had. Test administrators did a phone interview or a site visit to determine the level of existing infrastructure, and then used the appropriate delivery model for that school.

Internet delivery was used when the school had an appropriate computer lab with sufficient Internet connectivity and bandwidth. In this model, the school computers were used as the clients, with a remote server. This model requires minimal setup by the test administrator, and all data is collected centrally on the remote server, eliminating any data collection procedures required of the test administrator.

A carry-in server model was used at schools that had an appropriate computer lab with a local area network (LAN), but insufficient Internet connectivity to deliver the test via the Internet. For this model, test administrators travelled with a pre-setup server laptop that they plugged in to the school's LAN. The existing school computers were then used as clients to access the server over the LAN.

For schools that didn't have an appropriate computer lab, carry-in mini labs of computers were used (much like the CBAS model). The mini labs consisted of 10 computers; one for the test administrator (server) and nine student computers. The test administrator also carried in network hardware and cables. This model was only used for a handful of schools, usually quite remote and small schools.

Using this combination of delivery models depending on the school infrastructure, a very high success rate was achieved; 99% school-level response rate after replacements. Test administrator training was made more complex because each test administrator needed to be trained in all three delivery models.

Some technical issues with the ICTL study did cause some problems. The remote desktop client that exists on most Windows computers by default requires the use of a non-standard port, which is blocked by most firewalls in schools. The study used an Internet Explorer plug-in that allows remote desktop access through a standard port, but requires installation before the test session by the test administrator, and requires Microsoft Windows.

The future

At the moment there is no silver bullet for delivering computer-based assessment. The ideal technology for a study varies greatly depending on the objectives and requirements of the study. Simple, non-complex items may be delivered best using standard web technology available everywhere, whereas complex, rich items are best delivered using Flash or even remote desktop technology when application emulation is required.

The 'toolbox' of delivery methods used in the ICTL has a lot of benefits. The model is able to deliver very rich items, and a high rate of success is achieved by tailoring the delivery method to the school infrastructure. Having a choice of delivery models, along with the training involved, does raise the cost of the study, however.

In the future, Internet delivery should have the highest return on investment. Delivery through the Internet has the advantages of ease of deployment and low administration logistics and costs. The obvious disadvantage of Internet delivery at the moment when it comes to large-scale international studies is the lack of appropriate infrastructure.

The carry-in server model utilised in the ICTL study mentioned above is the best trade off at the current time. A high percentage of schools have a sufficient LAN but not sufficient Internet bandwidth to make Internet delivery possible. The carry-in server model has no need for any installation of software, and has the advantage of total control over the hardware and software on the server (as opposed to ERA where both server and client must run on unknown hardware).

References

Consortium (2004) CBAS Requirements and Specifications; Internal documents.

Consortium (2005) CBAS 2006 Preliminary Field Trial Analysis, Doc: CBAS(0510)2; Internal document.

Consortium (2006) CBAS 2006 Main Study Review; Internal document.

Consortium (2008) ERA Field Trial Reviews; Internal documents.

ACER (2008) MCEETYA ICTL Review Committee meeting record, Field Trial report.

ACER (2008), A Preliminary Proposal for an International Study of Computer and Information Literacy (ICILS), Prepared in response to a request from the IEA Secretariat for the 2008 IEA General Assembly.

The author:

Sam Haldane
Senior Software Engineer
Australian Council for Educational Research
19 Prospect Hill Rd, Camberwell
Victoria, Australia 3124
E-Mail: haldane@acer.edu.au

Sam Haldane is a senior software engineer with the Australian Council for Educational Research (ACER). He managed the technical aspect of the PISA 2006 CBAS study, overseeing the design and development of the software used, including authoring, viewing, translation and delivery systems. He is currently involved in the software development for the PISA 2009 Electronic Reading Assessment project being developed by ACER in collaboration with colleagues at the Public Research Centre Henri Tudor and the German Institute for International Educational Research (DIPF).

eInclusion, eAccessibility and Design for All Issues in the Context of European Computer-Based Assessment

Klaus Reich, University of Innsbruck, Austria
Christian Petter, Institute for Future Studies, Austria

Abstract
Computer-Based Assessment (CBA) is gaining more ground and is considered to bring a number of advantages compared to traditional paper-pencil testing. Although it is emphasised that CBA may also be beneficial to people with learning difficulties or people with disabilities, research specifically focusing on the area of eAccessibility in the context of CBA is rather restricted. Despite numerous initiatives on eInclusion or Design for All in general launched by the EU in the recent years, the area of CBA has not been explicitly targeted. Furthermore, even though several guidelines on CBA and accessibility have been issued, to the date no binding standards exist when it comes to making CBA accessible to people with disabilities.

Following the workshop carried out in November 2007 on “Quality Criteria for Computer-Based Assessment of Skills” van Lent (2008) asked for a research agenda on „developing best practices on scrutinizing performances of subgroups“. He especially pointed out opportunities and problems of computer adaptive tests in relation to the testing of individuals with special education needs. The question arises if there was a follow-up of his call for research. During the workshop held in Iceland in September 2008, unfortunately, little consideration was given to these aspects. eAccessibility and eInclusion issues were only rarely mentioned during the whole workshop. Furthermore, statements made were limited to the fact that computer-based assessment (CBA) „is not about testing reading capacity in Times New Roman 10pt“.

Starting from the fact that *Times New Roman* is a very inappropriate font to use in screen design due to bad screen readability, the authors of this chapter point to some more important aspects and methods to be taken into consideration in the development of ICT-based assessment: To what extent do legislation and policy development on eInclusion and eAccessibility in Europe provide guidelines that might be relevant for the field of CBA as well, and what role can methodological approaches, like design-for-all, play in developing computer-based testing environments. In that sense the paper narrows the field to the specific regulations and guidelines to be applied to CBAs and points out the positive effects of considering them from an early development stage onwards.

Opportunities and Challenges of Computer-Based Assessment (CBA)

The use of CBA as compared to paper-pencil testing is meant to bring advantages and added value especially in large scale assessments (Scheuermann & Pereira 2008). On the other hand, there are a number of aspects that need to be taken into account in the deployment of CBA. Scheuermann & Pereira (2008) mention in their introduction such aspects as software quality, secure delivery, reliable network, capacities, support, software costs for development and test delivery and licensing. Apart from these issues, however, one should not neglect the specific needs people with disabilities may have. Overall, people with disabilities constitute about 15% of the European population and many of them encounter barriers when using ICT products and services (European Commission 2005). Therefore, eAccessibility is relevant for a core group of some 84 million persons in Europe, 50 million of them in the age range 15-64 (European Commission 2008b, based on Eurostat data).

A number of authors, nevertheless, point to the positive effects implied by CBA when compared to traditional paper-pencil testing (e.g. Abell et al. 2004, Ball 2003). When talking about opportunities and advantages implied by CBA for disabled persons it appears necessary to take into account the needs and requirements of people with disabilities. Positive effects, as for instance observed by Ball 2003 or in ISO/IEC 23988:2007 include the following:

- „Learners with a cognitive disability or a lack of confidence may benefit from an assessed online discussion or online group work, as the speed of response is slower and the pressure of contributing to a face-to-face discussion is removed;

- Drag-and-drop, gap fill and multiple choice questions completed online can be easier for a person with mobility or visual impairment than trying to write by hand in a small given space;
- Submitting assignments through e-mail can be helpful for learners who have problems with mobility“ (Ball 2003, p. 4).
- „[...] some aspects of good design, such as clearly legible screens and consistent positioning of buttons, benefit all users not only those with disabilities“ (ISO/IEC 23988:2007, p. 17).

Certainly there are a lot more positive aspects of CBA for disabled persons. Furthermore, in this context it might also be interesting to observe positive effects for all people when more attention is given to the specific needs of people with disabilities, i.e. by implementing enhanced usability and accessibility features.

The need for accessible CBA becomes even more apparent considering the application of CBA in high stakes tests. There is a real danger to exclude certain groups of people from these tests by making them inaccessible or arguing for „special arrangements“ although there is no need for that. In contrast to certain statements that the consideration of eAccessibility issues might negatively influence results of these tests (cf. the statement cited in the introduction), the question has to be asked, how relevant these tests are, if they exclude certain user groups.

However, using technology in assessment procedures, while removing barriers for some learners, can potentially create other barriers, especially when using unfamiliar or poorly designed software packages (Ball 2003). Therefore, an approach to make CBA as widely accessible as possible would benefit a large group of users.

Design for All / Universal Design

Applying the methodology of *Design for All* (DfA) or *Universal Design* may help to avoid some of these problems from the beginning. DfA – a term used synonymously for Universal Design but more common in the USA - emerged out of architecture and has brought important results in industrial design and new media design. As Klironomos et al. (2005) state, it is now a well-defined body of knowledge addressing the design of interactive products, services and applications,

- which are suitable for most of the potential users without any modifications;
- which have standardized interfaces, capable of being accessed by specialised user interaction devices;
- which are easily adaptable to different users (e.g. by incorporating adaptable or customisable user interfaces).

The Center for Universal Design provides a concise definition of Universal Design: "*The design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design*" (http://www.design.ncsu.edu /cud/about_ud/ dprincipleshtmlformat.html#top). When applying the principles of *Design for All* to CBA forms of assessments are developed that "[...] *are designed from the beginning, and continually refined, to allow participation of the widest possible range of students, resulting in more valid inferences about performance*" (Thompson et al. 2005). As Thompson et al. (2005) put it, universally designed assessments are not intended to eliminate individualisation, but they may reduce the need for accommodations and various alternative assessments by eliminating access barriers associated with the tests themselves. The authors point out several elements of DfA, including an inclusive assessment population, precisely defined constructs, accessible, non-biased items, simple, clear and intuitive procedures, maximum readability and comprehensibility, amongst others. The consideration of *DfA* principles in the development processes of CBA right from the beginning may anticipate policy development to come in the field of eInclusion. In addition, the Information Society and Media Directorate (2005) of the European Commission has already recognised that „*the most cost-effective and non-discriminatory form of access to ICTs is through the Design-for-All process, where mainstream products and services are designed to be used by as many people as possible regardless of their age and ability*.“ Overall, developers of CBA should "*anticipate the variety of accessibility needs that may occur and seek to design in solutions to minimise the through life cost of accessibility*" (Ball 2006).

eInclusion Policy Development in the European Union

CBA that intends to be applied in Europe on a broader scale should consider the recent developments in the field of eInclusion as elaborated in a series of policy documents of European stakeholders in recent years. In these policy documents the general perception of an inclusive society is outlined, as well as means to achieve this. The European Commission launched the eEurope initiative in 2000 with the aim of accelerating Europe's transition towards a knowledge-based economy. In the 2002 Action Plan the adoption of the Web Accessibility Initiative (WAI) guidelines, the development of a European Design-for-All (DfA) curriculum and the strengthening of assistive technology and DfA standardisation were recommended (Council of the European Union and the Commission of the European Communities 2000). In line with this, the European Parliament in its 2002 Resolution on Web Accessibility, "*reiterates the need to avoid any form of exclusion from society and therefore from the information society, and calls for the integration of disabled and elderly people in particular*" (European Parliament 2002). Furthermore, the Ministerial Declaration on eInclusion proposed "taking all necessary actions towards an open, inclusive knowledge-based society accessible to all citizens" (Stephanidis 2003).

The eEurope 2005 Action Plan marked another step in aiming to mainstream eInclusion in all action lines. It also proposed the introduction of accessibility requirements for ICT in public procurement. In the same document the European Commission pointed out the benefits of a DfA strategy applied to products and services: „*DfA not only allows a more thorough consideration of accessibility requirements when designing a product or service, but also fosters important economies by avoiding costly redesign or technical fixes after their deployment*" (European Commission 2005, p. 7). In 2005 and the beginning of 2006 the European Commission published memoranda on overcoming the broadband gap and fostering support of eAccessibility. The Ministerial Declaration published at the conference in Riga (11.-13. June 2006) set the starting point for a European initiative for digital integration.

The i2010 Initiative is the EU policy framework for the information society and media. It points to the need to support inclusion, better public services and quality of life through the use of ICT (Commission of the European Communities 2005). The i2010 High Level Group has implemented an "eInclusion" subgroup in order to define the specific steps. The most recent cornerstone in this development process is the Ministerial Conference held in Vienna from the 30th November to the 2nd December 2008, including the ceremony of the eInclusion awards. On the 1st of December 2008 the Commission outlined the lack of coherence, unclear priority setting, and poor legislative and financial support as major aspects resulting in an insufficient impact of efforts in eAccessibility (Commission of the European Communities 2008a).

These EU policies have been reflected in national legislation by its member states to different degrees, i.e. several states have implemented anti-discrimination legislation and, to different degrees, have made eAccessibility mandatory e.g. for public websites (e.g. in Italy and Germany). The Commission itself states that "[...] there is considerable fragmentation in the treatment of eAccessibility, both in the issues addressed [...] and the completeness of policy instruments used" (Commission of the European Communities 2008a, p. 4).

Other countries like the USA (e.g. the Americans with Disabilities Act 1990) or Australia (Disability Discrimination Act) to some extent have a longer tradition in fighting discrimination.

eInclusion/eAccessibility in Existing Guidelines and Standards

In relation to assessment Ball et al. (2003) sum up the requirements "[…] that disabled learners must not be disadvantaged in education, and it is important to ensure that learners are not unfairly treated in assessment situations. Colleges have an obligation to anticipate the needs of learners and to make reasonable adjustments to ensure that disabled learners can demonstrate their skills and abilities equally with their non-disabled peers. This obligation extends to online, distance and blended learning" (Ball et al. 2003, p. 3). Even one year earlier Wilse (2002) asked for guidelines in relation to CBA and eAccessiblity, but remarkably, there is hardly any reference yet in policy documents that refer to the need of making CBA accessible by following available laws, guidelines and/or ISO norms. At the current status eAccessibility is recognised as an important aspect of CBA in different more or less informal guidelines (e.g. TestAccess or Accessible Assessments) but has not found the needed attention in binding standards and guidelines, which would be of special relevance for high stakes tests.

The British “Special Educational Needs and Disability Act (SENDA)” is one example that might lead the way for other policies at national and EU level to follow but, as Wiles (2002) observed, it does not elaborate specifically on assessing disabled students. By examining the ISO/IEC 23988:2007 some problematic issues can be outlined. Although the standard points to as different (and important!) aspects as usability issues, assistive technologies, alternative and enhanced output devices and giving additional times for test candidates using assistive technologies, etc., the standard lacks the following:

- the aim of the standard is to set out principles and good practice, but not the details of the means by which they are to be achieved;
- Usability standards are considered (e.g. ISO 9241), but there is a lack of accessibility guidelines (or references to the relevant standards/guidelines respectively);
- the need to use appropriate accessibility analysis tools to verify accessibility is recognised (ISO/IEC 23988:2007, p. 18), but the need for experts to review certain accessibility problems is not mentioned.

Interestingly usability standards are largely covered for instance in ISO/IEC 23988:2007 by providing references to the relevant standards, e.g. ISO 9241 - Ergonomic requirements for office work with visual display terminals (ISO/IEC 23988:2007, p. 3f), but there is an obvious lack of further specification of accessibility.

The Web Content Accessibility Guidelines (WCAG 1.0 and the new WCAG 2.0 Guidelines published in Dec. 2008) developed by the W3C (World Wide Web Consortium) provide a comprehensive set of criteria against which to test accessibility. In addition WCAG 2.0 provides a single shared standard and guidance for achieving Web content accessibility. Furthermore, automatic test tools use them as their reference. The Web Content Accessibility Guidelines are of relevance in other contexts as well, and different national adaptations, which might become relevant for CBA, use them as their basis, e.g. the German "Barrierefreie Informationstechnik-Verordnung" (BITV).
Besides these specific guidelines for web-based technologies, there is certainly a need to consider norms and guidelines in relation to operating systems and software in general. Here Microsoft Windows, Apple and the Linux community have developed standards, guidelines and software specifications to be considered.
ISO/TS 16071 - Ergonomics of human-system interaction - Guidance on accessibility for human-computer interfaces, paragraph 1194.21 of the Section 508 on „Software applications and operating systems" or software accessibility checklists by commercial institutions, e.g. the IBM Software Accessibility Checklist tackle software accessibility from different angles. The W3C User Agent Accessibility Guidelines (UAAG) address the accessibility of any software that retrieves and renders web content for users, e.g. web browsers, media players, plug-ins, and other programs, as well as assistive technologies.

However, the need for a consideration of certain eAccessibility requirements does not stop here but has to be reflected in other circumstances as well, e.g. in the field of item generation:

- Development of accessible items, cf. e.g. solutions proposed by Questionmark Computing Ltd. and work done by the Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES) resulting in the Test Accessibility and Modification Inventory (TAMI).
- Provision of an accessible environment for the development of these items. The World Wide Web Consortium in its Authoring Tools Accessibility Guidelines (ATAG) provides a profound framework for that.

In principle, guidelines and regulations on assessment as well as guidelines in the field of eAccessibility and the provision of ICT-based learning offers are of relevance for CBA. Many practitioners in the field might not be aware of this and the demands for guidelines in the field of CBA might be rooted in this fact. On a general basis, guidelines that already exist for the provision of ICT-based learning offers can be extended to CBA, since assessment may be a part of these learning offers. These guidelines, however, often provide very limited space for accessibility specifications and recommendations tend to be rather perfunctory. Therefore, claims for specific guidelines that take into account issues of accessibility in the context of CBA were already aired at the beginning of the 21st century for instance by Wiles (2002).

Conclusions and Recommendations

The consideration of accessibility aspects right from the beginning, following a Design-for-All approach, may not only make CBA accessible to a broader target group by also improving the usability and other aspects of the assessments, but allows avoiding high costs that may arise when systems have to be adapted in order to comply to certain standards at a later stage in their development process. This understanding follows the perception of inclusion by Abbott (2007) transferred to CBA. He argues that the needs of potential users are to be considered at an early stage and thus CBA should be set up to be inclusive, whether or not the need for such changes appears to be present.

However, the uptake of eAccessibility guidelines and standards, as well as Design-for-All methods by practitioners and researchers in the field of CBA is slow. Furthermore, existing standards in the field (e.g. ISO/IEC 23988:2007) remain on the level of recommendations, demanding for a more specific legislation especially when it comes to educational policy development. Policy has given very important impetus on different levels, in the field of CBA, however, there is certainly one missing.

The following two guiding principles of the code of practice for the use of information technology (IT) in the delivery of assessments (ISO/IEC 23988:2007, p. 7f) are of specific importance from an eAccessibility perspective:
"*a) […] using IT should not result in any reduction in the assessment validity or reliability; b) …delivery and scoring should be fair to all candidates, and as far as possible should not disadvantage any candidate as a result of factors which are irrelevant to the knowledge, understanding or skills being assessed*". In order to meet these principles, however, more work on different levels is needed, e.g. policy development, software development, training in eAccessibility and especially further extensive research on the accessibility of CBA. Therefore, following the statement of van Lent (2008) outlined in the introduction to this paper, we would like to further emphasise and refine this call. The following questions should be addressed as long as the development of CBA for high stakes tests is still in its initial development phase:

- Measurement of the effects of the consideration of eAccessibility guidelines and, if possible, of Design-for-All approaches in the development of CBA;
- Call for further discussion on ethical questions: in- /exclusion of specific groups of people from high stakes tests;
- Specific accommodations and to what extent they really have to be used when using Design-for-All approaches. To what extend do Design-for-All methods enable a broadening of the user groups for CBA?
- Further development of international standards and binding agreements for CBA, and/or the extension/adaptation of existing standards (e.g. ISO/IEC 23988:2007) in relation to the latest research in the field of eAccessbility respectively;
- Research on the results of tests carried out via paper-pencil tests, by providing special accommodations for disabled users and by using assistive technologies in tests that have been designed considering Design-for-All and eAccessibility principles.

We conclude this chapter with a statement by the European Commission (2005, p. 2) that sums up the basic perception of eInclusion to be uptaken by practitioners in the field of CBA as well: „The implications are clear: making the benefits of ICT available to the widest possible number of people is a social, ethical and political imperative.“ The present situation in the field, however, as well as in other fields, is still far away from this imperative.

References

Abell, M., Bauder, D., & Simmons, T. (2004): Universally Designed Online Assessment: Implications for the Future. URL: http://people.rit.edu/easi/itd/itdv10n1/abell.htm (last accessed, 21.11.2008)

Abbott, C. (2007): E-inclusion: Learning Difficulties and Digital Technologies. futurelab series, Report 15. URL: http://www.futurelab.org.uk/resources/documents/lit_reviews/Learning_Difficulties_Review.pdf (last accessed, 19.11.2008)

Ball, S. et al. (2003): Inclusive Learning and Teaching: ILT for Disabled Learners. 2003 URL: http://www.techdis.ac.uk/resources/files/Theme2.4.pdf (last accessed, 19.11.2008)

Ball, S. et al. (2006): Accessibility in e-Assessment Guidelines Final Report. Commissioned by TechDis for the E-Assessment Group and Accessible E-Assessment Forum. Report Prepared by Edexcel. 16th August 2006 URL: http://www.techdis.ac.uk/resources/files/Final%20report%20(TechDis)SBfinal.pdf (last accessed, 19.11.2008)

Commission of the European Communities (2005): Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. "i2010 – A European Information Society for growth and employment" {SEC(2005) 717}. Brussels, 1.6.2005 URL: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2005:0229:FIN:EN:PDF (last accessed, 19.11.2008)

Commission of the European Communities (2008a): Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. "Towards an accessible information society". COM(2008) [804] final. Brussels, 1.12.2008 URL: http://ec.europa.eu/information_society/activities/einclusion/docs/access/comm_2008/comm_en.doc (last accessed, 5.12.2008)

Commission of the European Communities (2008b): Commisson staff working document. Accompanying document to the Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. "Towards an accessible information society". Status and challenges of e-accessibility in Europe. COM(2008) [804] final. SEC(2008) 2916.Brussels, 1.12.2008 URL: http://ec.europa.eu/information_society/activities/einclusion/docs/access/comm_2008/staffwp.doc (last accessed, 5.12.2008)

Council of the European Union and the Commission of the European Communities (2000): eEurope 2002. An Information Society for All. Action Plan prepared by the Council and the European Commission for the Feira European Council. 19-20 June 2000. Brussels, 14.6.2000 URL: http://ec.europa.eu/information_society/eeurope/2002/documents/archiv_eEurope2002/actionplan_en.pdf (last accessed, 19.11.2008)

European Parliament (2002): European Parliament resolution on the Commission communication eEurope 2002: Accessibility of Public Web Sites and their Content (COM(2001) 529 - C5-0074/2002 – 2002/2032(COS)). URL: http://ec.europa.eu/information_society/policy/accessibility/z-techserv-web/com_wa2001/a_documents/ep_res_web_wai_2002.html (last accessed, 19.11.2008)

Information Society and Media Directorate (2005): An Information Society Open to All. Will information and communication technologies enrich and empower the lives of vulnerable groups or become an additional obstacle for them. (Factsheet, 12). URL: http://ec.europa.eu/information_society/doc/factsheets/012-eaccessibility.pdf (last accessed, 01.11.2008)

ISO/IEC 23988:2007 Code of practice for the use of information technology (IT) in the delivery of assessments, British Standards Institution. Formerly BS 7988:2002.

Klironomos, I., Antona, M., Basdekis, I., Stephanidis, C. (Eds.) (2005): White Paper: Promoting Design for All and e-Accessibility in Europe. URL: http://www.edean.org/Files/EDeAN_White_Paper_2005.doc (last accessed, 01.11.2008)

Lent, G.v. (2008): Important Considerations in e-Assessment: An Educational Measurement Perspective on Identifying Items for an European Research Agenda.In: Scheuermann, F. and Pereira, A. G. (Eds.): Towards a Research Agenda on Computer-Based Assessment. Challenges and needs for European Educational Measurement. Luxembourg, 2008. pp. 97 – 103

Ministerial declaration (2006): ICT for an Inclusive Society. Riga, Latvia, 11-13 June 2006. URL: http://ec.europa.eu/information_society/events/ict_riga_2006/doc/declaration_riga.pdf (last accessed, 19.11.2008)

Scheuermann, F. and Pereira, A.G. (Eds.) (2008): Towards a research agenda on Computer-based Assessment.Challenges and needs for European Educational Measurement. Luxembourg, 2008 URL: http://crell.jrc.ec.europa.eu/CBA/EU-Report-CBA.pdf (last accessed, 3.12.2008)

Stephanidis, C. (editor on behalf of the Greek Presidency) (2003): Ministerial Declaration on eInclusion. Ministerial Symposium „Towards an Inclusive Information Society in Europe“, organised jointly by the European Commission and the Greek Presidency of the Council, Crete, Greece, 11 April 2003. URL: http://www.eu2003.gr/en/articles/2003/4/11/2502/ (last accessed, 19.11.2008)

Thompson et al. (2005): Considerations for the Development and Review of Universally Designed Assessments. CEO Technical Report 42. URL: http://cehd.umn.edu/NCEO/OnlinePubs/Technical42.htm (last accessed, 3.12.2008)

Wiles, K. (2002): Accessibility and computer-based assesment. A whole new set of issues? In: Lawrie Phipps, Allan Sutherland and Jane Seale (Eds.) Access all Areas: Disability, technology and learning. London, 2002. pp. 61-65.URL:
URL:
http://www.techdis.ac.uk/resources/files/AAA.pdf (last accessed, 3.12.2008)

Web Resources:

Accessible Assessments, http://www.shu.ac.uk/services/lti/accessibleassessments/content/section_2/2.3.8.html

Americans with Disabilites Act 1990, http://www.ada.gov/pubs/ada.htm)

Apple accessibility, http://www.apple.com/at/accessibility

Australia Disability Discrimination Act 1992, http://www.austlii.edu.au/au/legis/cth/consol_act/dda1992264/

Barrierefreie Informationstechnik-Verordnung - BITV, http://bundesrecht.juris.de/bitv/index.html

Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES), http://peabody.vanderbilt.edu/x8312.xml

Gnome (Linux) accessibility, http://projects.gnome.org/accessibility

IBM Software Accessibility Checklist, http://www-03.ibm.com/able/guidelines/software/accesssoftware.html

Microsoft Windows Accessibility, http://www.microsoft.com/enable

Ministerial conference Vienna, 30th November - 2nd December 2008, http://www.bundeskanzleramt.at/site/4296/default.aspx

Questionmark Computing Ltd., http://www.questionmark.com/us/perception/section_508_compliance.pdf

Section 508, http://www.section508.gov

TestAccess, http://www.aph.org/tests/access/index.html

Test Accessibility and Modification Inventory (TAMI), http://peabody.vanderbilt.edu/TAMI.xml

W3C Authoring Tools Accessibility Guidelines - ATAG, http://www.w3.org/TR/WAI-AUTOOLS/

W3C User Agent Accessibility Guidelines – UAAG, http://www.w3.org/TR/WAI-USERAGENT/

W3C - WCAG 1.0, http://www.w3.org/TR/WCAG10

W3C - WCAG 2.0, http://www.w3.org/TR/WCAG20/

The authors:

Klaus Reich
University of Innsbruck
Innrain 52
6020 Innsbruck – Austria
E-Mail: klaus.reich@uibk.ac.at
WWW: www.uibk.ac.at

Christian Petter
Institute for Future Studies Austria
E-Mail: christian.petter@futurestudies.at
WWW: www.futurestudies.at

New Media/Teacher Training Coordinator (Vice-Rectorate Students/Teaching) and Project Officer (Institute of Educational Sciences), Klaus Reich specialises in continuing professional development for educators/trainers and in e-learning, including accessibility and usability of learning environments, barrier-free web-design, e-accessibility and evaluation of web-based learning.

Christian Petter works as a research associate and project officer for the Institute for Future Studies Austria, a non-profit institution of applied research. His current work field comprises - among other aspects of ICT-based learning - e-learning course development and research on ICT-based learning measures specifically targeting groups prone to be excluded from the information society like people 50+ or people with low educational achievement.

......III. Transition from Paper-and-Pencil Testing to Computer-based Testing

Risks and Benefits of CBT versus PBT in High-Stakes Testing

Introducing key concerns and decision making aspects for educational authorities

Gerben van Lent
ETS Global

Abstract
This paper aims to highlight the complexity of moving from paper-based to computer-based testing and introduces some key aspects of a decision making process that is based on a benefit-risk approach to facilitate this transition from paper-based to computer-based testing. It attempts to shed some light on the considerations that make test expert organizations like Educational Testing Service (ETS®) cautious in their advice and highlight some elements that can help educational authorities in their decision making. The paper limits its scope to high stakes testing, where high stakes is interpreted as test scores that have significant consequences for individuals e.g. certification exams, end of secondary graduation tests, admission tests.

There is a well known joke about a drunk having lost his car keys in the dark and looking for them under a streetlight, because there he is able to see. In discussing transitions to computer-based testing sometimes it seems to educational institutions asking testing experts for solutions that the latter behave like the drunk whereas the testing experts get the impression that the institutions randomly search in any dark place, just because that is the only certainty they have. I will start with some general considerations including some reasons why authorities want to transition from using one testing mode to another, followed by some recent research-based observations and conclusions, to illustrate some key issues and concerns.

A basic thought model is introduced consisting of an integrated decision making process and a benefit-risk analysis. Discussing a specific example and zooming in on a limited number of aspects instead of a holistic overview, serves to illustrate the complexity of managing such a transition both in the number of issues to consider as well as the level of difficulty. In the conclusions the aspects of accountability and the importance of communication are highlighted.

Few people these days will object to a conclusion that Randy Bennett from Educational Testing Service formulated in his May 2002 paper '*Inexorable and Inevitable: The continuing Story of Technology and Assessment*': the incorporation of technology into assessment is inevitable because as technology becomes intertwined with what and how students learn, the means we use to document achievement must keep pace'. Other issues raised in his paper might be seen as more controversial these days, e.g. that starting with multiple choice-based tests in low stakes testing situations might be advisable, because using a variety of more varied item types too quickly in high stakes environments has (too many) political, financial, legal but most of all educational risks. Bennett observes that cost elements, technical complexity, the need to maintain scale and a perceived sufficiency in making decisions based on multiple choice questions did slow down progress.

He ends on a positive note by stating: The question is no longer whether assessment must incorporate technology. It is how to do it responsibly, not only to preserve the validity, fairness, utility and credibility of the assessment enterprise but, even more so to enhance it.

Why computer-based testing is attractive for educational authorities

What are common reasons that educational authorities articulate in considering transitioning from paper-based to computer-based testing?

In England the QCA published their strategic decision in 2005 to promote e-assessment and to justify it by stating: E-assessment can provide timely feedback to inform future teaching and learning and 'when ready' assessments give learners greater ownership of their learning. With e-assessments it is possible to test areas of knowledge, skills and understanding that would be impossible using pen-and-paper-based testing.

Clearly here the educational benefits anticipate the closing of the gap between assessment and learning and specifically highlighted are the timely feedback, when-ready testing facilitated by increased access and more realistic/'appropriate' tasks.

In the Netherlands the CEVO (*Centraal Examencommissie Vaststelling Opgaven*) responsible for the end of secondary school graduation tests provides in their 2007 policy paper '*De computer bij de centrale examens – beleid voor de komende jaren'* the following reasons: Allows for testing of other skills thereby providing better coverage of the educational programs; better connection to what is required in continuation courses and programs; more attractive for candidates; more flexibility in test dates; and automated processes in administration, logistics and scoring.

In a paper that Julius Björnsson, from the Educational Testing Institute, Reykjavik, Iceland presented in 2007 at the workshop '*Quality Criteria for Computer-based Assessment of Skills'* organized by the Joint Research Centre of the European Commission, the advantages for Iceland compulsory education to transition their national tests to a computer adaptive format included: Shorter testing through adaptive tests allow for better test-student fit, quick results, better measurement of the extremes on both ends of the scale, less stress and press (not everyone at the same time), testing with modern technology which everybody is using all the time, more rich items and materials (multimedia), reuse of items, cheaper and quicker coding.

Lastly, a reason given recently by representatives of the Indian Institutes of Management, who are considering moving to computer-based testing, is that they believe it becomes easier to handle large numbers of test takers in a short period of time. This can be considered as a benefit of increased access. After all, to the extent that testing is distributed both geographically and temporally, access is theoretically improved. That said; however in the United States the two well known college admission tests, SAT and the ACT, test roughly 300,000 students on paper for admission into colleges at roughly 3000 test centers 6-8 times a year. This is not currently feasible with computer-based testing in the US and most likely not anywhere else. So handling of high volumes alone is in this example a weakness rather than a strength of computer-based testing.

Research findings about risks of computer-based testing

In the workshop that was organized in 2007 by the Joint Research Centre of the European Commission, a number of presentations referred to research that has been done in relation to computer-based testing. The picture that emerges from there as well as from ETS's continuous computer-based assessment research is that although many aspects have been scrutinized, especially where it relates to large scale testing and high stakes testing, there are still significant gaps in what we know or can claim with confidence, as illustrated below.

In his paper, "*Important Considerations in e-Assessment; An educational measurement perspective on identifying items for an European research Agenda"*, that is included in *Towards a Research Agenda on Computer-based Assessment* (2008) edited by Friedrich Scheuermann and Angela Guimarães Pereira which is part of the JRC Scientific and Technical Reports series, Van Lent concluded that, based on a number of research publications about comparability of scores and/or adaptive testing, many of the changes in computer-based testing seem to be driven first and foremost by what technology allows us to do. Educational research tries to catch up, but is lagging behind. For major high stakes testing programs this often leads to a situation that relatively 'old' methods of testing are used. The current state of CBT can be characterized as: some CBTs offer little or no added value; some "innovative" items are likely to contribute more 'artifactual' than valid measurement; limited site capacity often forces continuous administration, which can introduce serious security concerns; test administration algorithms are getting smarter but remain limited. Key issues that have to be taken care of are linked to: design, accessibility, and security. Appropriate underpinning with relevant research is an absolute necessity.

In the same publication (2008, p78) Oliver Wilhelm and Ulrich Schroeders cautioned: Obviously our current knowledge about the equivalence of assessment across test media is by no means sufficient to infer that complex measures can be used regardless of the test medium. It is desirable to clearly distinguish between changes in means and covariances in future studies investigating cross mode comparability. High or even perfect correlations between latent variables of a test administrated in more than one test medium are compatible with substantial changes in means. Therefore, comparisons across test media can privilege participants in one medium over participants in another medium even if the latent variables for the tests are perfectly correlated. Similarly the same test administrated in two test media might have the same mean and dispersion of scores but the two scores might have different reliability and the latent variables captured by both tests might not be perfectly correlated.

These remarks align with conclusions drawn by Edward. W. Wolfe and Jonathan. R. Manalo (2005, p51), in their research paper '*An Investigation of the Impact of Composition Medium on the Quality of Scores From the TOEFL® Writing Section:* A serious shortcoming of most research concerning score differences attributable to test delivery medium is the fact that most of these studies examine group differences rather than individual differences. These studies have suggested that, on average, there are only small differences between scores on computer-based and pencil-and-paper tests. Unfortunately, to our knowledge no studies have attempted to ascertain the magnitude of the impact of testing medium on individual examinees, particularly those who are members of groups who may be expected to be "at risk" due to lower levels of computer familiarity and comfort or higher levels of computer anxiety.

In their paper '*Challenges for Research in e-Assessment*', Jim Ridgway & Sean McCusker (p87 of the JRC publication) warn against overstating the chances of success in developing new measures that are ICT-based and work in the target audiences. They list a number of uncomfortable truths that impact the success of computer-based assessment including: Working with ICT across the educational sector is particularly difficult, because of the wide range of hardware and software platforms that are used; ICT has had very little impact on classroom practices – let alone on attainment; Optimistic claims for the likely effectiveness of e-assessment [especially e-portfolio work] are rarely grounded in evidence; such evidence as we have about the benefits of e-portfolios is weak, and discouraging; We know far too little about how to design assessment to support learning;

A balanced approach for educational authorities

It might look different so far, but the objective of this paper is not to argue against computer-based testing as such (although a decision could be made to wait or slow down the introduction of a computer-based testing solution), but to introduce a thought model for educational authorities for a risk-benefit approach to decision making regarding the change from paper-based testing to computer-based testing. The disappointing news is that there is no clear-cut solution that solves all uncertainties. The decision making process that is suggested is a heuristic benefit-risk model of thinking. It takes as a given that most technical difficulties (in terms of software and hardware) can be addressed although with possible substantial cost implications.

Conceptually when thinking about change the following model will support the decision making process:

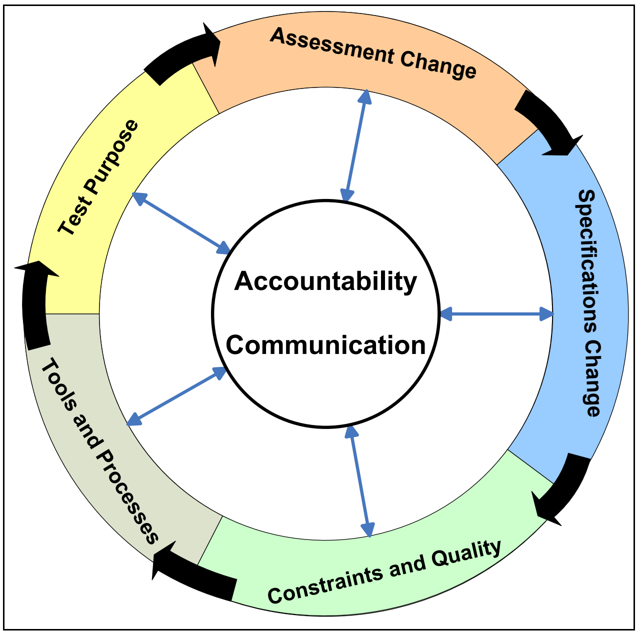


Figure 1: A free adaptation of a figure from Von Davier, Alina, (April 24th, 2008) Change and Stability in Educational Assessment: an Oxymoron?, Presentation held at Institute of Educational Assessors National Conference, UK

The starting point for transition is to articulate very clearly the existing purpose of the test. *Assessment Change* is the step where it is identified in how far there is the intent to change either the purpose and/or any of the core aspects of the designing, developing, delivering and reporting of a test as e.g. covered by ETS's Standards for Quality and Fairness. The next steps include defining and describing what the effects are of this change: How does it impact on the test specifications, which constraints do we have to take into account, what quality criteria are guiding the process, and what are the tools we want to use and the processes we need to implement in order to secure a controlled transition that preserves or improves the quality of the assessment as a whole?

In this paper where we explore the implications of transitioning from paper-based testing to computer-based testing, the phase *Assessment Change* can be characterized as follows: Driven by a political (stakeholder) decision to move towards computer-based testing under the condition that the overall purpose of the test has to remain the same (e.g. admission to higher education), the mode of testing needs to transition from paper-based to computer-based. One reason to discuss this particular choice is that while ideally the rationale ought to be that an articulated new test purpose leads to changes in the assessment process to optimize achieving the adapted test's purpose, in practise often changes in the assessment process are made externally under the constraint that the test purpose and all other aspects of the assessment are only minimally affected. One valid rationale for this type of decision could be the following: Few, if any existing tests, paper or computer-based, uniformly and ideally meet all requirements that their sponsors set down. This means that every testing program is a compromise that meets most of the requirements that the educational authorities consider important, but fails to meet other, presumably less critical requirements.

Importantly, computer-based testing and paper-based testing have quite different profiles of strengths and weaknesses. Until recently, educational authorities were forced to try to meet their program's requirements only through paper-based testing. So a valid and rationale reason for changing modes is that the profile of strengths and weaknesses associated with the now-available computer-based testing mode simply better meets the program's requirements (as weighted by the educational authorities' values).

In this particular case questions that arise are: What are the possible consequences of change for score meaning; How will that affect the decisions we want to make; and What can we do to preserve score meaning across changes so as to ensure validity, fairness and credibility? In other words preserving the accountability to the test taker for whom the scores have a significant impact on their future lives. The benchmark for the acceptable level of accountability and trust in high stakes testing is in this case the existing situation, where implicitly the assumption is made that the considerations to change are made in a relatively stable environment. This means that on average there is sufficient public trust in the current system and the certificates linked to the tests are accepted and perceived as having value. Although educational measurement experts and other educationalists might have their doubts, the system as is, functions de facto as a benchmark of quality.

Since we take the decision for mode change here as a given, the emphasis in exploring strengths and weaknesses is on *Specification Change and Constraints and Quality* as they relate to *Tools and Processes* that need to be made available and need to be implemented. In creating a transparent and easy to interpret profile, strengths have to be tied to weaknesses and this can be done best by linking benefits to risks. Sometimes we label this approach the 'no free lunch' axiom. The essential element here is that decision making is based on comparing profiles and not on disconnected lists of strengths and weaknesses. Of course the benefit risk model should ideally be applied before any decision regarding change has been taken, so that an option is to consider a profile 'not to change' or at least to consider more profiles.

In determining the adaptation to the test design when moving from one mode of delivery to another, many major and minor objectives (can) play a role, such as:

- Improve exam validity.
  - Develop high-quality items and test forms.
  - Standardize test administration conditions and improve exam proctoring.
  - Ensure scoring and reporting accuracy.
- Improve examinee access.
  - More administration dates.
  - More administration sites.

- Safeguard exam security:
  - Form security. Prevent unauthorized access to test forms prior to administration
  - Item security. Prevent examinees who have already tested from assisting those yet to test.

A wide variety of test designs can be considered for achieving these goals. However as indicated before, the goals sometimes conflict with one another, meaning no test design is likely to be ideal with respect to all, but we are looking for the optimal profile of strengths and weaknesses. Often the features that allow certain test designs to meet various goals are accompanied by associated costs or concerns. The challenge is to find the design that maximizes benefits while controlling potential risks.

The following table illustrates this balance between benefits and risks:

Benefit	Associated Risk
Improved access by increasing the number of test administration dates.	Use of the same test form across occasions entails security risks. Use of different forms across occasions both increases item development requirements and introduces the need to equate to make scores on alternate forms comparable.
Testing on computer allows introduction of dynamic and interactive item types that can improve measurement of existing constructs or allow measurement of new constructs.	By changing how an item is presented and appears, CBT administration can change the construct that an item measures.
Improve the quality of items and test forms by pretesting.	Introducing item pretesting complicates form assembly, production, and scoring. Item pretesting also entails a slight security risk.
Transition to online scoring of constructed response items by human markers.	Changing the media by which markers are trained, with which they score and how quality is assured can introduce different scoring behavior.

Figure 2: Benefit-Risk table

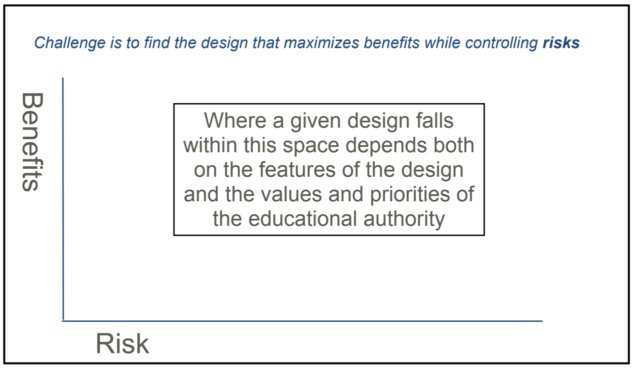
The associated risks in the table are examples of risks that fall in various categories including:

1. Unsustainable item development requirements.
2. Weak data collection designs that inadequately support necessary psychometric methods.
3. Complex operational logistics that invite error.
4. Likelihood of serious security breaches.
5. Limited test administration dates / sites that frustrate examinees and hold down volume.
6. Public perception as “behind the times”.

Risk management strategies would include classifying the risks under these headings, linking them to probability that they will occur and what the impact would be (translated into costs). Applying mitigation strategies would reduce the risk and therefore reduce overrunning the budget during implementation, but increase spending up front. The more complex a solution, the larger the investment will be.

The “best” design is that which most appropriately balances benefits and risks, where ‘most appropriately’ can be interpreted as fitting in the comfort zone of the educational authority. It is likely that authorities will look for solutions in the upper left corner of the box in the diagram below.

Specific example to illustrate complexity of decision making



Challenge is to find the design that maximizes benefits while controlling ***risks***

Where a given design falls within this space depends both on the features of the design and the values and priorities of the educational authority

Figure 3: The test design space

Let us explore this further through an example to illustrate the complex nature of the decision making and the benefit-risk assessment. Assume that a testing program with 300,000 test takers per year intends to move from paper-based testing in one sitting to computer-based testing. The decision has been taken and assessment experts are asked to provide advice how to best move forward. The option to continue (an improved form of) the current paper-based process is politically not acceptable. However, all other aspects of the assessment process including the scores are seen as quality benchmarks that cannot be compromised (also not retrospectively). The purpose of the test (selection for admission) stays the same. It is also assumed that the quality principle that test scores are valid and fair regardless of where or when a candidate takes the test is adhered to. Other quality aspects and constraints include:

- Equal candidate access
- Maintain and where possible improve validity:
  - High-quality items and tests
  - Standardized testing conditions
  - Accurate scoring
- Secure exams:
  - Physical security
  - Item security

In discussing *Constraints and Quality* aspects, a key issue that arises is: Can the computer-based testing take place in one session (capacity) or not? If not this leads to questions like how many sessions are needed, is there a preference for continuous testing or during a specific time window on fixed dates, how are the current forms treated (confidential or published), etc.

Moving from one session to more sessions means automatically that more forms and more items are needed, that testing circumstances need to be comparable across sessions and that scores from one session have the same meaning as scores in another session within one test cycle (not taking into account if there is any formal equating or some other process of maintaining standards in current testing procedures from cycle to cycle).

In considering multiple forms, two approaches seem possible to secure comparable results. There are many variants of these approaches; the goal here is simply to give a general sense. The approaches are as follows:

1. Build multiple forms and spiral all or some of them randomly in all administrations. Use common population equating approaches.
2. Build multiple forms with variable sections to be used for equating/scaling. This approach could allow for exposing a form in only certain administrations.

The second approach is maybe better but more complex, harder to get right, and more expensive.

How would the first approach work? Assume 10 forms worth of items would be developed. You might then assemble e.g. 30 forms; individual items or sets would appear in 3 forms. The forms would be designed as parallel as possible, although this will be based on judgment. In the most extreme version of this plan, you could randomly spiral these 30 forms during the entire testing window. From a security perspective, candidates would have a one-in-ten chance of seeing an individual item, and a one-in-thirty chance of seeing an intact form. By the end of the testing window, each form will have been taken by 10,000 candidates (this is more than sufficient for equating purposes) spread over e.g. 60 administration days (with 2 sessions per day) if we assume a total test centre capacity of 2500. So, key factors of the risk profile are that you expose all the items in the pool from day one, but then spread the risk over all administrations.

There are, of course, security risks with this plan. The security profile could be improved by raising the level of item development. For example, if you developed 15 forms worth of items each item would appear in only two forms, and the chances of seeing an individual item go down to one in fifteen. At an extreme, developing 30 forms worth of items could lead to no overlap between forms. However, the increased item development will tax schedules and quality, and of course lead to increased expense.

The second approach could work as follows: Again 10 forms worth of items would be built and assembled in 30 forms. In addition, external equating anchors would be built that might be a variable section in the test. Because of common items, you would not need to rely on spiralling, and in fact would want all the people in an administration to get the same form (to make sure there is a sufficiently large equating sample) on a test administration day. Thus, when you assume you can test as much as 2500 student per session, an individual form would show up in four sessions (so exposed again to 10,000 students in groups of 2500 at a time) i.e. on two test administration days (of 60 in total), and an individual item in 12 sessions.

As with the first approach, the security characteristics of this system could be improved by increasing item development assuming test centre capacity is fixed. Key factors of the risk profile are that you expose one test at a time, and it is then exposed one more time to the whole testing population on that second day.

So, with the numbers given above, there is already an initial increase by a factor of 10 in item development. What implications does that have on the timeline for item production, the number of item writers during the transition, and costs? In addition to item writing, review procedures would require more people, thereby creating a need for standardization, or require the existing staff to spend more time on this activity.

What are the implications for the maintenance of the pool of items, including the refreshment rates in general, but also the refreshment rates per item type or item aspect (difficulty level, certain content coverage)?

Maintenance of the item pool is closely linked to the security aspects of the transition to a multi test environment. It touches an essential risk, namely that a test taker has an unfair advantage over other test takers due to unauthorized prior knowledge of the test items.

How does the risk arise?

There are two cases:

1. When the test or constituent items are stolen prior to administration.
2. When the test or constituent items are seen during an administration, remembered, and communicated between test takers who are taking the test or items at different times.

When does the risk arise?

Case 1: Theft prior to administration: The stealing of tests or items can occur any time from the point of item conception to the finalization of the actual test form. The impact of the theft is greatest, of course, when the test is complete.

Case 2: Communication of test or item content between administrations of the test or items: this case occurs because the specific test form or some numbers of its constituent items are administered repeatedly across occasions. A single administration strategy or computer-based strategy that uses a new form at each test administration would eliminate this risk. However, the move to computer-based testing and the multiple administrations per cycle assumed in this example, lead to the reuse of some number of items and tests.

How can the risk be mitigated?

Case 1: Theft prior to administration:

This risk can be well managed through the rigorous human, physical, and electronic security infrastructure and protocols that have been developed over decades of development, printing, and distribution of high-stakes paper-based and computer-delivered tests. The methods employed by professional organizations like ETS®, Prometric, the awarding bodies in the UK, CITO, just to name a few are examples of organizations that have well-developed techniques for maintaining test security. Mature security procedures (at a cost) make this type of security breach rare.

Case 2: Communication of test or item content between administrations of the test or items.

The risk that a test taker has benefited unfairly from prior knowledge can be lessened, but not eliminated, by:

- minimizing the reuse of items and/or tests
- disrupting the business or cooperative model of organized cheaters
- decreasing the ability of test takers to predict when specific items or tests will appear

An important question requiring analysis concerns the size and structure of the item pools and related implications for pool security. Although pool security is impossible to evaluate precisely or guarantee in practice, it is commonly measured by the extent of overlap or shared item content across papers given to randomly selected pairs of examinees. By this measure, all testing programs that pretest or equate generate a positive overlap. In their paper '*Improving Security under Continuous Testing*', Tim Davey and Elizabeth Stone (April 2007) first warn against the danger not so much of items and item pools being disclosed as such but that they are disclosed so quickly that economics and logistics make it impossible to develop replacement items and pools fast enough to keep up with item loss. At that point a testing program becomes either invalid or unviable.

The theme of their paper is to promote and discuss techniques that testing organizations can use to stave off this prognosis. They outline the mechanisms that examinee populations and coaching schools can use to reconstruct, distribute, learn and recall test content followed by actions that testing organizations can take to counter, disrupt and minimize each of these activities.

Conclusions

In this paper I have tried to highlight a number of issues that educational authorities and decision makers should be aware of when considering to move paper-based high stakes testing programs to computer delivery. Research shows, illustrated here with a limited number of examples, that there are a significant number of issues that need attention, and that there should be awareness that some might even not be satisfactorily resolved for the time being. This might seem in contradiction with what often seems to be the perception today, that the transition to computer-based testing is more a technical and logistical issue than an educational measurement issue. An important reason for this 'lighter' attitude might be that there is not sufficient awareness among policy makers and educators of the implications of high stakes testing with respect to individual performance as opposed to group performance.

The conference, for which this paper was written, is entitled '*Lessons learned from the PISA 2006 Computer-based Assessment of Science (CBAS) and implications for large scale testing'*. Although the impact of PISA is significant for countries and educational systems it isn't for individual test takers, and therefore doesn't qualify as a high stakes test in the description used here. This means that on the one hand my observations wouldn't necessarily apply to PISA and other sample-based assessments, but conversely neither would observations based on sample-based testing necessarily apply to high stakes testing.

From a fairness and validity point of view it is not defensible for individual test takers to be disadvantaged because of a change in medium that results in the possible introduction of construct irrelevant variance. For educational authorities it is neither defensible from an accountability perspective nor from a legal perspective, the latter growing in importance with the increasing public attitude towards making use of complaint procedures including litigation.

Instead of looking for a solution that is full proof, I introduced a decision making model that helps identify key issues that have to be taken into account. This decision making model advocates to create different profiles how change can be implemented. The starting point should always be the articulation of the purpose of the test and the definition of what is going to be changed, either in support of achieving that purpose or as an externally imposed change that necessitates action to preserve achieving the original purpose of the test. Then, through introducing a benefit-risk analysis, the paper highlights some choices that can be considered, and their resource and cost implications. In making a final decision the profiles could be visualized in a benefit-risk diagram that includes also a 'comfort zone' as defined by the decision maker in which a solution must lie in order to qualify.

All elements of the process are linked through the overarching perspective of accountability to the test taker. Transparency is advocated through the central role of communication. From a communication perspective not only test takers are important, but a range of stakeholders including test takers, parents, teachers, school administrators, score users (employers, institutes of further education, etc.), public at large, government and the media.

Information needs to be provided amongst others about:

- The change itself
  - Why change from PBT to CBT?
  - Will the new CBT test be better?
  - Was there something wrong with the PBT test?
- The effects on the learning process
  - How will the change affect teaching?
  - How do we prepare the test takers?
- The quality of the new test and the information gathered
  - How will the change affect Reliability, Validity?
  - Is multi-year trend data important?
  - Will degree of computer proficiency affect test taker performance?
  - Will the scale change?
  - How do we use the scores?
  - What are the implications for score-user systems and databases?

This information can only be provided if the 5 elements of the decision model (*Define Test Purpose, Assessment Change, Specifications Change, Quality and Constraints, Tools and Processes*) have been thoroughly addressed and there is a clear view of associated benefits and risks. Communication itself is also an essential element of mitigation strategies for many associated risks.

Finally: Educational Authorities have the responsibility to implement assessment policies and practices that serve the needs of the learners and meet the requirements that society places on education. Researchers and expert organizations have the obligation to support them in meeting this challenge by discovering and developing cost-effectively designed systems of assessment that address these needs, in other words: 'no drunks, no darkness, no disillusionment'.

Acknowledgement

I thank my ETS® colleagues Tim Davey, senior research scientist, and John Dumont, Director of Educational Solutions who coordinated the transition of ETS's Test of English as a Foreign Language (TOEFL®) from paper-based testing to computer-based testing, for their contributions to this paper and their helpful comments on an earlier draft.

References

ETS Standards for Quality and Fairness http://www.ets.org/Media/About_ETS/pdf/standards.pdf

Bennett, Randy Elliot (2002) Inexorable and Inevitable: The Continuing Story of Technology and Assessment Report Number: RM-02-03, available from http://www.ets.org/research/researcher/RM-02-03.html

CEVO;(2007) De computer bij de centrale examens-beleid voor de komende jaren, available (in Dutch only) from http://www.cevo.nl/9334000/d/
cevo-comp_bij_ex.pdf:

QCA Factsheet (2005) QCA leading the way in e-assessment, available from http://www.qca.org.uk/libraryAssets/media/6929_factsheet_e-assessment.pdf

Ridgway, Jim & McCusker, Sean (2007) Challenges for Research in e-Assessment', available from http://crell.jrc.ec.europa.eu/CBA/EU-Report-CBA.pdf.

Van Lent, Gerben (2007) Important Considerations in e-Assessment; An educational measurement perspective on identifying items for an European Research Agenda, available from http://crell.jrc.ec.europa.eu/CBA/EU-Report-CBA.pdf

Wilhelm Oliver & Schroeders Ulrich (2007) Computerized Ability Measurement: Some substantive Do's and Don'ts, available from http://crell.jrc.ec.europa.eu/CBA/EU-Report-CBA.pdf

Wolfe, Edward. W. and Manalo, Jonathan. R. (2005) An Investigation of the Impact of Composition Medium on the Quality of Scores From the TOEFL® Writing Section: y, available from http://www.ets.org/research/researcher/RR-04-29.html\

Davey Tim, Stone Elizabeth (April, 2007) Improving Security under Continuous Testing, 2007, Paper presented at the annual meeting of the National Council of Measurement on Education Chicago

Von Davier, Alina, (April 24th, 2008) Change and Stability in Educational Assessment: an Oxymoron?, Presentation held at Institute of Educational Assessors National Conference, UK

The author:

Gerben van Lent
ETS Global
Strawinskylaan 913
1077 XX Amsterdam
The Netherlands
Tel: +31 (0)20 880 4161
E-Mail: gvanlent@etsglobal.org

Gerben van Lent has primary responsibility for directing the development of new business for ETS Global. This covers all or a subset of the core areas of expertise of ETS: test design and development, test delivery and scoring, test analysis and reporting or research and development. Van Lent represents ETS Global and ETS in dealing with potential clients, government officials, other external organizations, agencies and at professional meetings. He has presented or conducted workshops at international assessment conferences about quality assurance, large scale assessments and assessment and technology issues, and has contributed articles to journals in the field of education and measurement. He is an external member of the Research Committee of AQA in the UK and a board member of ACETS, the Arabic Centre for Educational Testing Service.

Transformational Computer-based Testing

Martin Ripley

Abstract

This article contrasts two approaches to the use of technology to support assessment and testing. A migratory approach is described, whereby test providers seek to move traditional paper-based tests to screen-versions. This approach can bring administrative gains and service improvements. However, the approach is contrasted with a transformational approach, whereby the test developer sets out to redefine assessment and testing approaches in order to lead educational change. The article argues that, far from being expensive and unreliable, innovative test developers have proven the educational value of assessments which test skills and knowledge in the context of the 21st century. The article suggests that an increasing focus on transformational approaches is warranted.

Even the most cursory survey of current attitudes to e-assessment suggests that not everyone is convinced by the argument that computer-based testing can – or should – lead to educational change. Tom Burkard at *The Centre for Policy Studies* in England says that multiple choice questions are more accurate than essays in assessing students' writing. Essays written by students in England nowadays are often "virtually unintelligible" with even basic errors not being corrected, he claims. Tom Burkard's report points to the potential to save significant sums of public money by replacing essays with multiple choice exams. He also argues that it is no longer possible to find enough teachers to mark tests accurately – which rather begs a question of how those same teachers manage to mark their own students' classwork.

For other commentators, a trend toward multiple-choice questions represents a significant concern that technology will drive us to more straightforward forms of testing, making these ever more routine and predictable. In 2006 one of the UK's largest awarding bodies, the *Assessment and Qualifications Alliance* (AQA), completed its first trial of computer-delivered assessment at GCSE, England's main examination for 16 year-old students. The approach taken by AQA was to create a closed-response computer-delivered test as one component of a science GCSE. The on-screen test was created by selecting suitable materials from past paper-based tests. This AQA pilot was critically reviewed by the national media, who were sceptical of the value of multiple-choice testing. Jonathan Osborne, Professor of Science Education at King's College London, said: "How is this going to assess pupils' ability to express themselves in scientific language, a major aspect of science?" The Times newspaper expressed strong doubts regarding the educational value of this approach to testing, a view shared by many educators in the UK (The Times Online, 2006).

More encouragingly, in a third development, Cisco, Intel and Microsoft launched a Call to Action in London in January 2009. (Cisco 2009) The purpose of this call to action paper is to advocate the case for innovation in assessment and learning. The paper argues strongly that high stakes educational assessments and tests define the outcomes expected of schools and teachers, and for there to be significant reform of learning in the 21st century there must first be significant reform of assessment. "Reform is particularly needed in education assessment—how it is that education and society more generally measure the competencies and skills that are needed for productive, creative workers and citizens. Accountability is an important component of education reform. But more often than not, accountability efforts have measured what is easiest to measure, rather than what is most important. Existing models of assessment typically fail to measure the skills, knowledge, attitudes and characteristics of self-directed and collaborative learning that are increasingly important for our global economy and fast changing world. New assessments are required that measure these skills and provide information needed by students, teachers, parents, administrators, and policymakers to improve learning and support systemic education reform." (Cisco 2009, p2)

Defining e-assessment

For the purposes of this chapter, a broad definition of e-assessment is needed. (JISC 2006)

- E-assessment refers to the use of technology to digitise, make more efficient, redesign or transform assessments and tests;
- Assessment includes the requirements of school, higher education and professional examinations, qualifications, certifications and school tests, classroom assessment and assessment for learning;
- The focus of e-assessment might be any of the participants within the assessment processes – the learners, the teachers and tutors, managers, assessment and test providers and examiners.

Overview

This chapter seeks to answer the charge that the use of technology in assessment inevitably leads to the erosion of educational standards, and the concern that the computer cannot enhance the ability of the human teacher or marker to interpret complex evidence of students' test performances.

An overview of current trends and developments in the use of technology in assessment might well conclude that it (technology) is being used in the most straightforward, objective and multiple-choice forms of assessment. This chapter seeks to highlight the ways in which technology can be used to design unprecedented assessments which enhance learning and which reflect well the priorities of nations in designing educational systems for the 21st century.

An e-assessment model

Advocates of e-assessment frequently point to the efficiency benefits and gains that can be realised. These benefits might be to do with the costs of test production, the ability to re-use items extensively, to create power- and adaptive-tests, or to build system improvements, such as test administrations systems which are able to provide tests when students want. However, advocates of e-assessment less frequently look to the potential of technology to support educational innovation and the development of 21st century skills, such as problem solving, communication, team working, creativity and innovation.

The following diagram provides a representation of the contrast between these two drivers: the business efficiency gains versus the educational transformation gains.

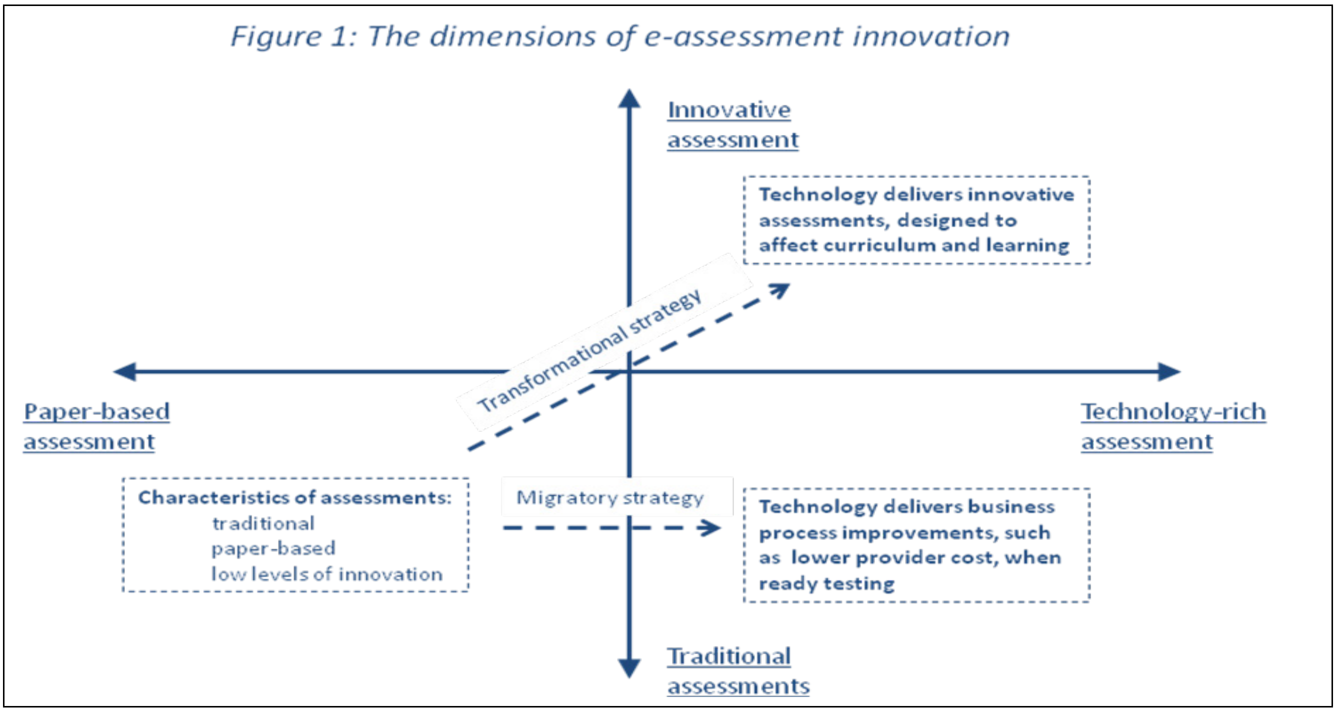


Figure 1: The dimensions of e-assessment innovation
Innovative assessment
Technology delivers innovative assessments, designed to affect curriculum and learning
Transformational strategy
Paper-based assessment
Technology-rich assessment
Characteristics of assessments: traditional paper-based low levels of innovation
Migratory strategy
Technology delivers business process improvements, such as lower provider cost, when ready testing
Traditional assessments

Figure 1: The dimensions of e-assessment innovation

The diagram contains four quadrants, three of which are of relevance here.

The lower-left quadrant represents traditional assessments, typically paper-based and which do not exhibit any tendency to develop or innovate year-on-year. The bulk of school-based and college-based assessment and tests reflect these characteristics.

Moving from the lower-left to the lower right quadrant represents a migratory strategy. Paper-based assessments are migrated to a screen-based environment, but are level qualitatively unchanged. One example would be to migrate a paper-based multiple-choice test to a screen-based test. The example given above of an AQA project to migrate some science test content to science is an instance of this approach.

The lower-right quadrant not only represents tests which have been migrated from paper. It also includes computer-based tests and assessments which have been developed from scratch but which, nonetheless, reflect closely many of the characteristics of traditional paper-based tests. There are benefits of a migratory strategy and there are compelling reasons for placing straightforward tests on screen. These benefits include:

- Providing tests at a time of a student's choosing – or, when ready testing.
- Reducing costs.
- Improving test reports and diagnostic analysis of students' performance.
- Improving marking reliability.
- Speeding up the marking and reporting cycle.

However, the concerning characteristic of e-assessments in the lower-right quadrant is that there is no innovative effective on the curriculum, teaching or learning.

By contrast, the upper-right quadrant represents a transformational strategy underpinning the use of technology in assessment. The defining characteristic of innovative assessment is that it is designed to influence (or minimally to reflect) innovation in curriculum design and learning. For example, a computer-based assessment of problem solving – which has used technology to innovate and redesign the nature of the problem-solving domain – seeks to provide an assessment of skills and abilities not normally assessed through paper-based tests.

Although there are few instances of transformative e-assessment, the projects that do exist provide us with a compelling case for researching and investing in assessments of this type. In 2005, Ken Boston, then Chief Executive of the UK government's curriculum and assessment regulator, spoke optimistically of a forthcoming transformation of assessment in which technology was presented as a catalyst for educational change. "*There is much less risk, and immensely greater gain, in pursuing strategies based on transformational onscreen testing: transformational question items and tasks; total learning portfolio management; process-based marking; and life-long learner access to systematic and personal data. There is no political downside in evaluating skills and knowledge not possible with existing pencil-and-paper tests, nor in establishing a new time series of performance targets against which to report them*"(Boston, 2005).

The main characteristic of transformative assessments is that the question items and tasks are transformed. The paper published in 2009 by Cisco, Intel and Microsoft - "*Transforming Education: Assessing and Teaching 21st Century Skills*" – sought to provide an illustrative example of this type of assessment. They described a plausible description of the type of task that might be assigned to a student completing an ICT test. "*Students are given a problem scenario in which they are rangers for a national park in which there has been a dramatic increase in the population of hares that threatens the ecology of the park. They are given the task of deciding whether or not to introduce more lynx into the system and, if so, how many. Students receive, respond to, and initiate communications with other rangers who are working on the project and have specialized knowledge of the situation. They search the World Wide Web to find out pertinent information on both hares and lynxes. They organize and analyze this information and evaluate its quality. They make predictions based on their analyses, test their predictions with modelling software, and analyze the results, as represented in graphs, tables, and charts. They integrate these findings with information from other sources and create a multimedia presentation in which they make and defend their recommendations and communicate these to others (Example courtesy of Edys Quellmalz)*"(Cisco2009, p14).

There already exist examples of this approach to test design. In England, a test of ICT skills was commissioned by the government as early as 2000. (See Boyle 2005 and 2006.) The purpose of the tests was to test the ICT skills of 14 year-olds. That project led to the development of extended, authentic tasks assigned to students completing tests of ICT skills in a virtual desktop environment. Students logged into the test environment and were presented with tasks to complete. Working in the virtual environment, in one task students were asked to create a job vacancies page for the local virtual newspaper. To complete this task, students had to research job vacancies across the myriad websites within the virtual environment, collating information, sending out virtual e-mails to clarify and confirm details; students were expected to respond to e-mails from the newspaper's virtual editor, requesting updates on progress. In other words, the assessment task was designed – successfully – to reflect real-life tasks and to present students with an authentic, simulated environment within which to complete the assigned task.

Towards an agenda for building transformational assessments

This example of transformative assessment is today an aspirational vision. It requires not just the design of authentic tasks, assessed within simulated environment, but also a root-and-branch transformation of all aspects of the testing process. To run the ICT job vacancies assessment described above, for example, it is also necessary to rethink and create robust solutions to the following aspects of tests.

Accessibility arrangements

The ways in which the test is designed to provide equal access for all learners, regardless of any special needs. In the domain of traditional, paper-based tests, most countries have developed sophisticated processes for ensuring access – whether through permitting amanuenses, extra time, Braille versions or many other measures. To be successful, the transformational assessments must also invent new technology based accessibility processes and procedures.

The equivalence of standards

Tests are designed to award grades to students, and in the late twentieth-century many governments around the world use those test results to measure trends and improvements in educational outcomes. In the context of this usage of test results, it becomes increasing incumbent on test providers to demonstrate – year-on-year – that the test results are being maintained. In this context there are emerging two schools of thought regarding the potential effects of introducing e-assessment into large-scale, high stakes test programmes.

- That providers of computer-based tests should ensure that standards of equivalence are maintained with paper-based antecedents (It should be borne in mind that assessment researchers have cask very considerable doubt on the ability of paper-based test providers to maintain standards over a period of many years. The seminal standards over time study led by Alf Massey led to a conclusion that five years is about the maximum length of time that any government should seek to draw standards-based comparisons of trends and improvements. (Massey 2003).
- That new time-series comparisons should be started afresh, marking the beginning of computer-based tests.

The former of these two attitudes is a risk-averse position to adopt, and leads test providers to limit and control innovation. The second position – which was clearly advocated by Ken Boston above – permits and encourages innovation.

Scoring and marking

One of the greatest challenges for the developers of transformative assessments is to design new, robust, comprehensible and publicly acceptable means of scoring student's work. Neither classic test theory nor IRT is sufficiently adaptable to reflect a measurement of students' performances in completing the job vacancies research task outlined above. That assessment task needs to be designed to reward the students according to the ICT skills they demonstrate in completing the task – it is not acceptable to collect the students' final newspaper job vacancies pages, and to surmise and estimate which ICT skills might have been used. The assessment requires the development of agreed problem solving heuristics and the development of models which can articulate clearly the differences between more- and less-eloquent solutions.

We are today far from the creation of these models. However, there are projects which have designed new approaches to measurement, to reflect the transformed design of the test. For example, Goldsmiths College in London has been working for five years on the design of a 6-hour practical task-based assessment. (Kimbell, 2006) Students are required to narrative a multi-media digital record of their practical activity, producing a journal of what was done. The assessment requires students to reflect, to record, to summarise, justify and to explain their solution to the practical task. And at the end of the assessment, the students' digital portfolios are collected and marked remotely by trained human markers. The approach to scoring that is being developed is based on Thurstone's graded pairs. The markers are presented with two students' portfolios at one time, and are asked to make a judgement about which is the better.

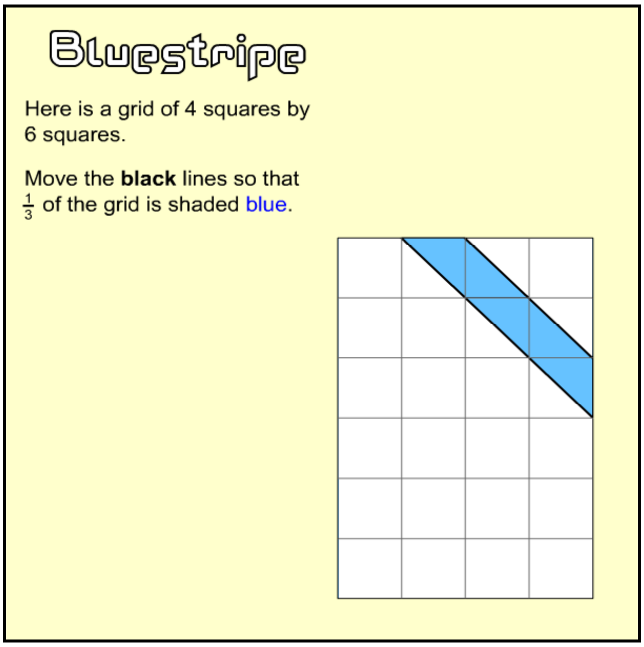
That is the only judgement markers make about the pair of portfolios. They do not score or grade at all. Once the first judgement has been made, the marker is presented with a second pairing, then a third, and so on.

After two large-scale trials of this approach to scoring, it has been found each portfolio has been compared to about 17 other portfolios, quite remarkably high levels of reliability have been achieved. *"[The reliability coefficient] value obtained was 0.95, which is very high...[This can be compared with] the Verbal and Mathematical Reasoning tests that were developed for 11+ and similar purposes from about 1926 to 1976. These were designed to achieve very high reliability, by minimising inter-marker variability and maximising the homogeneity of the items that made up [the test]. KR20 statistics for those were always between 0.94 and 0.96. ... [The Goldsmiths' test] has achieved this without reducing the test to a series of objective items*." (eSCAPE Phase 3 report, not yet published.) The Goldsmiths' project has demonstrated the feasibility of designed radically different measurement models, with sacrificing human input or reliability.

Describing new skills domains

By definition, many transformative assessments will operate in non-traditional skills domains – such as problem solving, or team working, communication or innovation. One set of tests that has sought to create and develop a new domain are the World Class Tests developed in the UK. These designed were inspired by the UK government. The design brief was to create on-computer problem-solving assessments for highly able 8-13 year olds from around the world. The World Class Tests cover a number of problem-solving domains. Peter Pool led the development of mathematical problem-solving World Class Tests at the University of Leeds. He describes the brief for the tests. "*The assessments are not about seeing how much mathematics has been covered - the questions do not require knowledge of mathematical content beyond normal expectations for students at ages 9 and 13, so acceleration through the curriculum brings no great advantage. The questions are about how deeply the mathematics is understood and they offer success to those who can bring insight, perseverance and flexibility of thought to a question*." (Pool 2006, p2)

One example of a World Class Tests item is *Bluestripe*. The following description is taken from Peter Pool's paper. This is a grid of squares with an adjustable shaded band. Each of the two slant edges of the shaded area can be moved parallel to its starting position.



Bluestripe

Here is a grid of 4 squares by 6 squares.

Move the **black** lines so that $\frac{1}{3}$ of the grid is shaded blue.

Figure 2: Example of a World Class Tests Question - Bluestripe

Moving either or both of the parallel sides can cause the shaded area to change its shape from trapezium to pentagon to hexagon to parallelogram. Its area needs to be 8 squares (from the information in the question), but it is difficult to operationalise this fact in any formulaic way over such a wide range of shapes, though whole and half squares are easily countable in individual cases. There are a number of approaches available - from counting squares to using the formula for the area of a parallelogram or trapezium, (knowing that the diagonal of a square is√ 2 x side length). But none of these is likely to be deployed before some exploration has taken place using the interactivity of the diagram. This allows students to see the shapes that are possible, to recognise those that have areas that are easy to calculate, to get a sense of an approximate answer or to notice other aspects that might suggest a way forward. Theoretically, there is an infinite number of possible solutions though most would require precise measurements that are not possible for the student on a computer screen - itself an additional factor for the student to take into account. Three of the more likely solutions are:

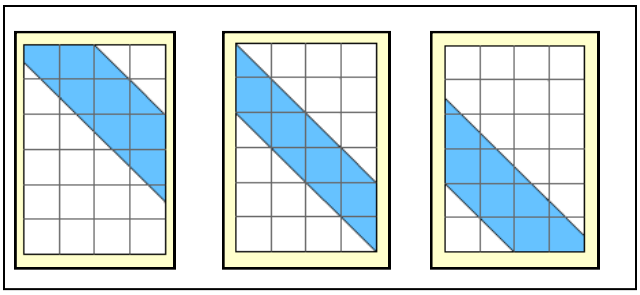


Figure 3: World Class Test Bluestripe solutions

In the middle diagram the shaded band can be seen as having four identical vertical parallelogram sections, each one square wide and having a structure of one whole square and two half squares. In the left and right hand diagrams the small shaded part squares can be matched to small white sections to make complete shaded squares.

None of these solutions requires advanced understanding of how to calculate areas of shapes. What is more useful is the ability take advantage of the interactivity to recognise useful features that can be investigated and from which a strategy can be evolved. It is worth noting that it would be almost impossible to ask this question on paper in such a way that a student would be confident she had understood the procedure; the practicalities of then doing the question on paper raise further issues of manageability. In this sense, the mathematics here is 'new' in so far as it would not (or could not) be presented in a conventional paper assessed curriculum, though the question itself remains very accessible to anyone who understands conventional mathematics.

Research completed by Valsa Koshy and Ron Casey indicates clearly that World Class Tests do assess problem solving skills not traditionally assessed. (Koshy 2001) They found that students who have highly sophisticated and well-developed problem-solving strategies perform well on World Class Tests. However, they also found that the same students did not perform as well on traditional test of mathematics. Koshy and Casey developed the term "submerged talent" to capture the notion of World Class Tests capability to identify latent problem-solving talent at a very high level of refinement. Arguably young children showing the skills required for high performance on World Class Tests are already evidencing the types of skills required in the 21st century.

Technological issues

Developing and implementing large-scale transformational assessment requires solutions to significant technological barriers and problems. No school system in the UK yet has the configuration of equipment, with staff trained and systems established to undertake complex simulation-based assessments. Parents and politicians will rightly worry about the potential for cheating or loss of students' data through insecure IT systems. Even if not based on evidence, many will harbour concerns that IT projects always run over budget and run late. These infrastructural issues today remain a significant set of barriers.

Demand

A final word has to be left to the issue of demand. There is still today, in 2009, scant evidence of demand for transformational assessments. Few teachers, few parents and very few students are calling for this approach to assessment reform. It is possible that the lack of demand for assessment reform is in part fed by perfectly natural wishes of parents and teachers to prepare students to perform well in today's tests – the tests that will determine entry to college and early careers. In this context, it would be a highly risky strategy for any high school principal to adopt innovative approaches to testing, without evidence of the acceptable currency of those tests alongside clear knowledge that teacher s are ready and trained to support students in the learning programmes implied by the innovative tests.

Some conclusions

The argument for transformative assessment is that it can act as a catalyst for significant educational change and can deliver a better alignment with the needs of 21st century learners. It is difficult to be optimistic that current trends are moving our assessment systems in the right direction. Most e-assessment implementations are non-transformational, and our policy makers are at best too uncertain and cautious to sanction large-scale transformation approaches. In England it took around 25 years from the introduction of calculators on a wide-scale until they were first expected to be used in school examinations. In 1994 and 1995 examination setters began to set mathematics tests which required students to use a calculator (alongside second papers which prohibited their use). Within a very few months, mathematics teachers began to teach students the skills of using a calculator. A 25-year gestation period for the calculator does not augur well for more radical innovative assessments.

References

Boston, Ken (2005) 'Strategy, technology and assessment', speech delivered to the Tenth Annual Round Table Conference, Melbourne, October. Online, available at www.qca.org.uk/qca_8581.aspx

Boyle, Andrew (2006) An evaluation of the decision to base the key stage 3 ICT test on a bespoke virtual desktop environment http://www.naa.org.uk/libraryAssets/media/ks3ictreportboyle.pdf

Boyle, Andrew (2005) Sophisticated Tasks in e-assessment: What are they? And what are their benefits? Online, available at http://www.caaconference.com/pastConferences/2005/proceedings/BoyleA2.pdf

Cisco, Intel and Microsoft (2009) Transforming Education: Assessing and Teaching 21st Century Skills http://www.latwf.org/docs/Transformative_Assessment--A_Call_to_Action_and_Action.pdf (Accessed 21st January 2009)

JISC (2006) eAssessment Glossary – extended. Online, available at http://www.jisc.ac.uk/media/documents/themes/elearning/eassess_glossary_extendedv101.pdf (accessed 23rd December 2008)

Kimbell, Richard (2006) eSCAPE Phase 2 and a number of other reports on Project eSCAPE are available online at http://www.goldsmiths.ac.uk/teru/projectinfo.php?projectName=projectescape

Koshy, V.and Casey. R (2001) Submerged Talent and World Class recognition in Assessing Gifted and Talented Children in Assessing Gifted and Talented Students (ed Carolyn Richardson) London: QCA 2002

Massey, Alf et al (2003) Comparability of national test standards between 1996 and 2001. QCA, London 2003.

Pool, Peter (2006) Losing your inhibitions: possible effects on assessment of dynamic, interactive computer items. Paper presented at the IAEA conference 2006, Singapore.

Times Online (2006) Select one from four for a Science GCSE by Tony Halpin. Online, available at: www.timesonline.co.uk/article/0,,22219509,00.html

The author:

Martin Ripley
3 Hampstead West
224 Iverson Road
West Hampstead
London NW6 2HX
E-Mail: martin.ripley1@btinternet.com

Martin Ripley is a leading international adviser on 21st century education and technology. He is cofounder of the 21st Century Learning Alliance, and owner of World Class Arena Limited. He is currently working with a number of public sector organisations and private sector companies in Asia, Europe and the USA. In 2000 Martin was selected to head the eStrategy Unit at England's Qualifications and Curriculum Authority (QCA). He won widespread support for his national Vision and Blueprint for e-assessment. He led the development of one of the world's most innovative tests – a test of ICT for 14 year-old students. Martin has spent 15 years in test development. He has been at the heart of innovation in the design of tests in England: he developed England's national assessment record for 5 year-old children; he developed England's compulsory testing in mathematics and science for 11 year-olds; he introduced the UK's first national, on-demand testing programme; to critical acclaim, he developed World Class Tests - on-screen problem solving tests that are now used world-wide as a screening tool for gifted students. These problem solving tests are now sold commercially around the world, including in China and Hong Kong.

Reflections on Paper-and-Pencil Tests to eAssessments: Narrow and Broadband Paths to 21st century Challenges

Katerina Kikis-Papadakis & Andreas Kollias
IACM/FORTH and Panteion University, Greece

Abstract
The present paper discusses some of the main issues that have been raised in the discourse on the use of ICTs for assessment in formal education and training. A main point is that this discourse is often grounded on a "pragmatic" view which approaches the use of ICTs in assessment processes from the perspective of massive assessment mechanisms which have been established during the 20th century. From this perspective e-assessment is often understood as migration from paper-and-pencil testing to ICTs-based testing. However, such a perspective may create the ground for systematic exclusion or disadvantage of certain subgroups of learners who may not have a fair opportunity to practice with ICTs and become familiar with the conditions and procedures in e-assessments. It is proposed that the discourse on e-assessment should be re-contextualized within a wider dialogue among all stakeholders in education and training about what are the things we want to achieve with the use of ICTs for learning. Such a dialogue should be supported by research in real-life, "ordinary", schools which would focus on the design and implementation of ICTs-related innovations organically embedded into pilot curricula and attainment targets and experiment with different formative and summative assessment techniques.

Within EU one emerging stream of discourse about the use of ICTs in education, training, lifelong learning and professional development is focusing on the potentials of implementing ICTs in assessing the knowledge and skills of young people and adults. A pragmatic view is often embedded, implicitly or explicitly, in the discourse under way. This takes as a given the fact that throughout the world and at various levels of formal education/training systems there have been established massive summative assessment mechanisms with the primary purpose to measure the knowledge and skills of thousands and sometimes millions of examinees.

In the Netherlands, for example, each year around 200 thousand students in their final year of secondary education take part in national examinations (Martinot, 2008, p. 49). In Greece every year more than 120 thousand Lyceum graduates participate in national exams leading to a place in tertiary education. This year around 170 thousand people took a (multiple choice) test which is necessary to get access to a job in the public sector. Furthermore, tens of thousands of university graduates are now preparing to take exams which lead to a limited number of new school-teacher posts offered by the Ministry of Education.

The reality described above is quite common around the world. Such summative assessments can affect in a very definite way the academic and professional future of examinees because test scores are often the most crucial data on which businesses, educational institutions, and governments base their decisions about recruiting, hiring, and promoting the most qualified candidates. No individual can really ignore the specific demands made by assessment mechanisms without undermining his/her chances to fulfill his/her aspirations for academic and professional progress. Similarly, teachers cannot ignore these demands no matter how much they believe in the validity of the measurements or in the teaching and learning style and practices that are most effective for the specific testing context and situation. In preparing their students they have to simply "teach to the test". In any other case they may risk the academic and professional future of their students and their own professional reputation and prospects.

These formative assessment mechanisms are historically the product of the huge expansion of education and training systems during the second half of the 20th century which was driven by the growing demands for skilled workforce in the advanced and developing economies around the world. However, paper-and-pencil tests, the landmark of massive assessments of the industrial epoch, become less and less relevant and cost-effective in information societies. One, rather popular, solution is the use of ICTs in assessment processes in ways that do not require major changes in the ways massive assessment mechanisms operated in the paper-and-pencil era and, furthermore, respond to their demands in efficient and effective ways. Some of these mechanisms have already adopted this approach by migrating, partially or wholly, from paper-and-pencil to screen testing. For example, in Europe, each year around 15 thousand people take the Graduate Management Admission Test (GMAT) (GMAC, 2008, p. 4), which is partly delivered in the form of computer-based adaptive multiple choice items. Other similar examples are the Test of English as a Foreign Language (TOEFL) and the Test of English for International Communication (TOEIC) that offer Internet-based tests which are taken by hundreds of thousands of people around the world.

Main issues related to e-assessment

The discourse around the comparative (dis)advantages and novel challenges raised by the migration to electronic assessment focuses mainly on the following broad issues:

a) economic: such as the investments required for the setup, operation and support of reliable systems for electronically administered tests (hardware, software and human resources) and the cost-effectiveness of ICTs-based testing as compared to traditional testing,
b) technology-related: such as user registration, authentication and management, interoperability of testing systems, test security, test-generation and delivery algorithms, the development and validation of technical standards etc,
c) test-related: reliability issues which have to do with possible differences in test results that may be due to the testing mode and test validity issues which arise from possible discrepancies between what is measured by paper-and-pencil and screen tests. More specific issues have to do with the reusability of test items, the adaptability of test items' level of difficulty based on prior responses, the enrichment of electronic items with multimedia materials (video clips, animations etc) and with new item types which demand from examinees to interact with digital objects such as simulations (and the related complexities involved in the automated measurement of user performance), etc,
d) electronic testing conditions and procedures: assessments based on paper-and-pencil tests follow strict regulations and rules which are aimed to ensure that the examination conditions and procedures are exactly the same for all examinees. The use of ICTs in examinations introduce important issues of potential variability in testing conditions and procedures, such as hardware and software specifications, internet connection speeds etc which have to be addressed, and
e) human recourses issues, focusing mainly around the training of the people who will be involved in the administration of screen tests in test-centers.

Advantages of e-assessment

The advantages of migrating from paper-and-pencil to ICTs-based testing are more or less similar to those identified for a wide range of activities that were previously conducted with the use of analogue materials and by hand.. Arguments in favour of e-assessment as compared to paper-and-pencil often stress on advantages such as:

a) faster administration, processing and delivery of test results to examination bodies and examinees,
b) error-free marking of true/false items,
c) readily available data for further statistical analysis,
d) enhanced interactivity and items which are composed of multimedia objects, allowing for the measurement of skills not easily measurable by traditional tests,
e) more flexible opportunities for self-practice "on demand" (provided that the examinees and their teachers are offered access to a test-generation and delivery sub-system of the electronic testing system),
f) more flexible assessment delivery which offers more opportunities for "assessment on demand", and
g) radical cuts in waste paper that is generated by paper-and-pencil tests.

Despite, however, the advantages of ICTs-based over paper-and-pencil tests, replacing paper-and-pencil tests by screen tests in real life, especially "high stakes", assessments which involve large number of examinees is easier said than done.

Challenges in the implementation of e-assessment in formal education and training

One of the most challenging issues for the assessment bodies is to ensure that all examinees face the same testing conditions and procedures. Among other things, this requires that all examinees have access to a strictly pre-specified set of hardware, software and related infrastructure in approved test centers. One option would be to develop screen tests which have system and hardware requirements that can be satisfied even by relatively outdated computers and which are experienced through a common to all user-interface, irrespectively of the variability in the capabilities of the available ICTs in test centers. Understandably enough, ICTs-based tests which are required to work equally well in outdated and modern computers are difficult to exploit most of the capabilities of today's technologies and therefore they offer rather limited added-value as compared to traditional tests. On the other side, the development of tests which run on up-to-date ICTs would require huge initial investments and sustainable funding policies for the development of advanced test items to measure skills that are difficult to measure by paper-and-pencil tests, the creation of large item databases, the implementation of adaptive test sequences or the operation and maintenance of test delivery servers which can satisfy the huge real-time demands of on-line test administration to thousands of examinees (see, for example, Luecht, 2005).

Furthermore, the replacement of paper-and-pencil by screen tests in "high stakes" exams unavoidably introduces equal opportunities' issues which are of paramount importance, particularly in public formal education and training (see AERA, APA, NCME, 1999, p. 74). Despite the growing availability of ICTs and broadband internet connections at home and at schools, even the most advanced countries in the world are faced with tough digital divide issues within their populations. In the USA, for example, only one fifth of the low-income Americans enjoy broadband connections. In contrast, 85 % of the upper-income Americans have access to such services (Horrigan, 2008, p. 3). The latest data show that in EU (27), around 40 % of the households are still not connected to the Internet (Eurostat, 2008, p. 2). Furthermore, only 30 % of people aged 16 to 74 report that they have more than "low" level basic Internet skills (i.e. knowing, for example, to do something more than using a search engine or e-mailing with an attachment) (Source: Eurostat, Information Society Statistics, 2007 data). The above indicate that the use of ICTs-based tests may systematically exclude or place in a disadvantaged position certain subgroups of learners who may not have a fair opportunity to practice with the tests and become familiar with the testing conditions. For those students who do not have access to computers or internet connection at home, it is more difficult to familiarize with ICTs because at schools regular access to them is often not easy. According, for example, to a relatively recent survey, in only around 12 % of primary and secondary schools in France the computer: student ratio is 1:5 or lower. In Germany this ratio was reported in only around 10 % of the schools surveyed, while in Italy in around 5 % (Benchmarking Access and Use of ICT in European Schools, 2006, p. 67).

Another factor that may place certain groups of examinees in a disadvantaged position in formal education and training is the practices and attitudes of the teacher population towards the use of ICTs in schools. While few teachers would be expected to have negative attitudes towards learning activities based on analogue materials such as textbooks, the use of computers for learning is not always a welcomed option. It is characteristic, for example, that despite the growing presence and use of ICTs in schools, one fifth of primary and secondary teachers in EU (25) tend to believe that computers do not have significant learning benefits for their students (Benchmarking Access and Use of ICT in European Schools, 2006, p. 375).

Furthermore, among teachers of the same country and subject matter large variations can be observed in the degree to which they use ICTs during their lessons. As shown, for example, on figure I below, 8.2 % of Humanities and Social Sciences teachers in France (2006 data) reported that they made use of ICTs in more than half of the total lesson time. On the other side, 25.7 % of their colleagues reported that they used ICTs for teaching no more than 5 % of their lesson time. Understandably enough, in case they had to participate in an ICTs-based assessment, the students of "ICTs-free" humanities and social sciences lessons would be in a disadvantaged position as compared to their peers in lessons with very frequent use of ICTs.

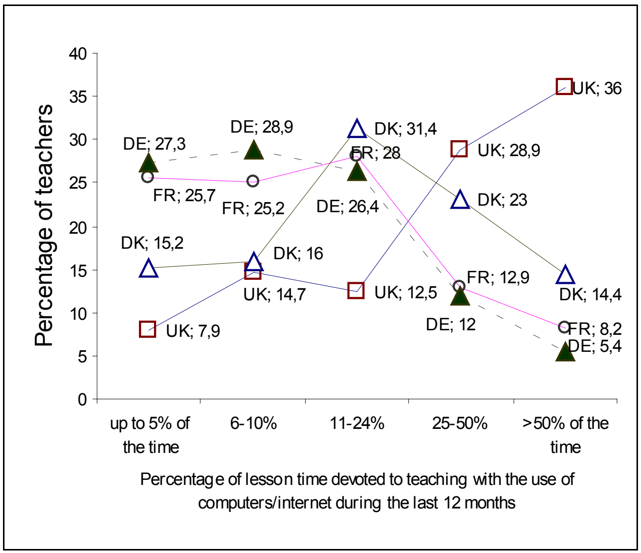


Figure 1: Percentages of Humanities and Social Sciences teachers (Y axis) grouped by country and percent of lesson time devoted to teaching with the use of computers/internet in class (Source: Benchmarking Access and Use of ICT in European Schools 2006, data obtained from tables 5.25, 5.28, 5.32, 5.36 & 5.40, pp. 189-209).

Intra-national variations in the use of ICTs for teaching are also observed in Science, Mathematics & Computer Sciences lessons. This variation is represented in the figure below.

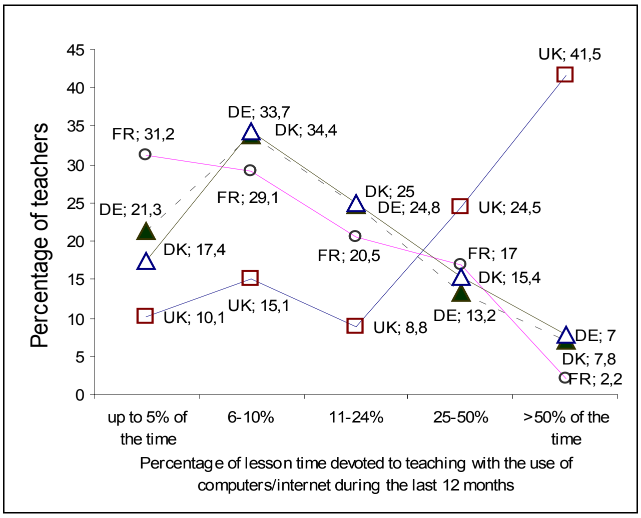


Figure 2: Percentages of Science, Mathematics & Computer Sciences teachers (Y axis) grouped by country and percent of lesson time devoted to teaching with the use of computers/internet in class (Source: Benchmarking Access and Use of ICT in European Schools 2006.

The data presented on figure II above show that in UK, one of the leading countries in the world in the integration of ICTs in everyday school activities, more than 25 % of those who teach Science, Mathematics & Computer Sciences lessons reported that they did not devote more than 10 % of their yearly lesson time teaching with the use of computers. On the other side, a sizable 41,5 % used computers during more than 50 % of their lesson time. Such kinds of discrepancies in the use of ICTs in everyday classroom teaching and learning could very likely create a ground for unfairness in "high-stakes" mandatory e-assessments against students who only occasionally experience the use of ICTs during their everyday maths or science lessons.

E-assessment within a wider dialogue and research framework on curricula and attainment targets

The discussion above suggests that e-assessments do offer self-evident solutions which would allow to the massive assessment mechanisms of the 20th century to continue doing their job in a "business as usual" style into the 21st century. On the other side, despite the calls about the need for radical reforms in curricula which would help young people develop lifelong learning skills and the widespread adoption of a socio-cultural and constructivist pedagogic rhetoric from individual teachers to Ministries of Education and powerful international corporations, there is little hope that ICTs can in effect become drivers for change in the ways our societies understand and practice "assessment" for academic and professional purposes. This is because school curricula and centrally defined attainment targets in formal education and training only occasionally presuppose explicitly the use of ICTs as *sine qua non* for teaching and learning (Kollias and Kikis, 2005).

The dominant assumption is that knowledge and skills are largely independent from the ways we learn and the tools we use for developing new knowledge, skills and attitudes. Therefore, the message that is communicated to the world of education and training is that ICTs is rather an add-on to its functioning which at best can "enhance" academic achievement and future professional prospects than change how, what and when we learn and hence change the ways we approach the assessment of new knowledge and skills that result from this process. The discussion about assessment with the use of ICTs has to be re-contextualized into a wider dialogue among all stakeholders in education and training about what are the goals and objectives we wish to achieve with the use of ICTs for learning. Such a dialogue should be supported by research projects in real-life, “ordinary”, schools which would focus on the design and implementation of ICTs-related innovations organically embedded into pilot curricula and attainment targets and experiment with different formative and summative assessment techniques. Such research projects should be extended in time so that they can produce some grounded evidence on what new curriculum areas, attainment targets and assessment techniques can be adopted by educational authorities on a wider scale and how change can be achieved in ways that are manageable by the schools and the educational actors.

References

AERA, APA, NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education) (1999) Standards for educational and psychological tests. Washington, DC: American Educational Research Association.

Benchmarking Access and Use of ICT in European Schools 2006 (2006, August) Final Report from Head Teacher and Classroom Teacher Surveys in 27 European Countries. http://ec.europa.eu/information_society/eeurope/i2010/docs/studies/final_report_3.pdf.

Eurostat (2008) Internet usage in 2008 – Households and Individuals, Data in Focus, 46.
GMAC (Graduate Management Admission Council) (2008) European Geographic Trend Report for GMAT Examinees 2003-2007. http://www.gmac.com/NR/rdonlyres/EA9C7488-6F94-412D-9C72-9221D76AF860/0/EuropeanGeoTrend07Web.pdf.

Horrigan, J.B. (2008) Home Broadband Adoption 2008. Pew/Internet. http://www.pewinternet.org/pdfs/PIP_Broadband_2008.pdf

Kollias, A. and Kikis, K. (2005) Pedagogic Innovations with the use of ICTs, from wider visions and policy reforms to school culture. Barcelona: Universitat de Barcelona.

Luecht, R.M. (2005, April) Some Useful Cost-Benefit Criteria for Evaluating Computer-based Test Delivery Models and Systems, Journal of Applied Testing Technology, 7(2). http://www.testpublishers.org/jattabs.htm..

Martinot, M.J. (2008) Examinations in Dutch secondary education - Experiences with CitoTester as a platform for Computer-based testing, in F. Scheuermann & A. G. Pereira (Eds), Towards a Research Agenda on Computer-Based Assessment: Challenges and Needs for European Educational Measurement. Institute for the Protection and Security of the Citizen, Joint Research Centre, European Commission.

The authors:

Contact person:
Kathy Kikis-Papadakis
Foundation for Research&Technology-Hellas
Institute of Applied Mathematics
PO Box 1385
71110, Heraklion, Crete, Greece

katerina@iacm.forth.gr

Dr Kathy Kikis is in-charge of the Educational Research and Evaluation Group of IACM/FORTH, has co-ordinated a number of EU-funded research projects focusing on innovation in education from a socio-cultural and organizational perspective.

Dr Andreas Kollias is member of the Educational Research and Evaluation Group of IACM/FORTH and has participated in several EU-funded projects in the area of e-learning and ICTs innovations in education and training. During the last five years he also offers research methodology courses at the Department of Political Science and History, Panteion University in Athens.

Transition to Computer-based Assessment
Motivations and considerations

René Meijer
University of Derby

Summary
Making the transition to computer-based assessment is a complex decision. It is inevitably riddled with a web of complex effects and dependencies. This article attempts to create an overview of some of the key considerations around validity that need to be included in such a decision. The overview is based on a series of workshops on this subject in Reykjavik, Iceland in September 2008. One of the prominent pitfalls is to treat this transition as simply a substitution of one instrument for another. The risk in this approach is twofold. Firstly, the transition inevitably has wider implications as each instrument of measurement also influences the system it measures in. As such changing an instrument inevitable changes the system in which it operates. Secondly, a substitution strategy will miss important opportunities to improve the overall value, validity and reliability of the assessment strategy (or even risk decreasing its validity). This is particularly important where signs of dissatisfaction with current assessment practice are already prominent.

Concurrency

The motivations to implement computer-based assessment (CBA) are varied, but have often been rooted in a drive for increased efficiency. From that perspective the logical question when evaluating the merits of CBA becomes: "Can we produce the similar, or better, results with less effort". It is against this background that initially discussions on validity are often framed in the narrow context of concurrent validity: "How will a computer-based assessment compare to our current assessment". Questions around gender equality for instance tend to be inspired by perceived differences in the comparison of outcomes of paper based exams with those obtained through CBA. While it is important to recognise and research these differences, it is equally important to investigate their true underlying causes, which might not be directly related to the medium at all. Sometimes this can lead to a new understanding of cognitive differences between different groups of people (sexes, cultures, socio-economic backgrounds, etc.), instead of warranting the invalidation of a valuable instrument. An example of this can be found in the work of Sørensen (2008), who stresses that there are many factors alongside gender, for instance culture and interests that influence assessment. She also raises that it might not be the medium of technology that in itself causes gender or other biases. Instead these differences might be more inspired by other changes that accompanied the transition to Computer-based Assessment, such as the reading-load of items, and question context presented in them.

The observer effect

Assessments do not function in isolation. They are part of a pedagogical and socio-political system of education. And while they are intended to measure and observe what goes on in this system, by doing so they also inevitably influence it. In experimental research this is called the observer effect, which states that the act of observation inevitably changes the phenomenon that is being observed, as a result of the observer becoming a part of the system, and thus interacting with it.

One way in which assessments feed back into the system, is by their value implications (see figure 1). The fact that something is assessed, not assessed, or even how it is assessed, conveys a message of what is important within a given domain. Sometimes this feedback emphases or de-emphasis the importance of content within the domain, but it can also have an impact on the type of proficiency learners try to achieve, or even what they perceive as proficient. When we analyse outcomes of tests in algebra, geometry and calculus, we might find that we can produce the same outcomes by just using the outcomes of the tests in geometry and calculus. However, this might lead to students neglecting their algebra, as it not assessed. And so, while this elimination was valid based on past data, it loses that validity with the implementation.

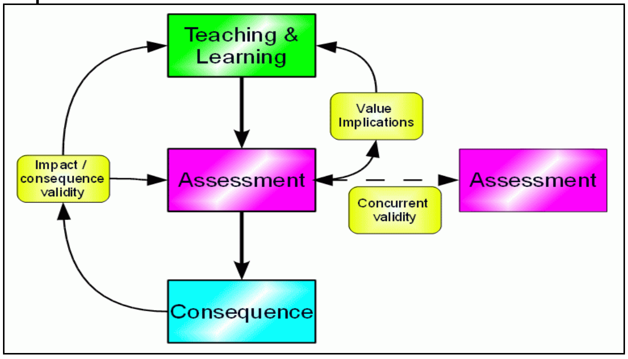


Figure 1: Assessment in context

This feedback effect that assessment has on learning should be taken into account whenever we chose to assess a construct through a new proxy. A good example can be found in the electronic marking of essays. Most electronic essay marking systems analyse essays on linguistic characteristics, the proxy, and not on content. While the results of this analysis are very comparable to those of a human marker, the consequences of this way of marking should not be underestimated. As explained by Bridgeman (2008) "the machine cannot evaluate the quality of an argument. A long, grammatical essay may receive a high score from a machine even if the argument is fallacious". And so these systems can be fooled, and could assign very high scores to an essay which in terms of content has no value, but adheres to the right linguistic and grammatical criteria. The implication of this, aside from the obvious risks around cheating, could be that form is perceived to be more important than substance.

Equally important is the recognition that the actions taken based on the outcome of an assessment also greatly influence validity (see figure 1). They have an impact on the assessment itself, as behaviour during the assessment changes based on the perceived consequences of the outcome. Getting a good understanding of someone's weaknesses and misconceptions in a high-stakes exam will be difficult, as the candidate has a vested interest in trying to hide these. This tension makes it problematic to combine formative and summative assessments. But the impact can reach much further. The Commons schools, children and families committee in the UK was quoted by the BBC (2008) in saying: "over-emphasis on their results can distort how children are taught" and "children's access to a balanced education is being compromised". Because the consequences in terms of funding and publicity of the SAT tests in the UK are so significant, they risk becoming more important than learning itself. This is one of the reasons why *Skolestyrelsen* (the Danish national school agency) made a conscious decision to not publish the results of the national tests in Denmark for rankings (Wandall, 2008).

Authenticity and design

There are other risks linked to the focus on concurrent validity. The comparison made when evaluating concurrent validity is somewhat narrow in that it focuses on the instrument that is the assessment. But it is important to realise that an assessment is more than just an instrument, but instead is a process intended to allow conclusions to be drawn in relation to a (set of) predefined question(s). This process starts with the formulation of a question, based on which an instrument is designed. This instrument is used to collect the data based on which a conclusion with regards to the question can be drawn (see the figure 2 below).

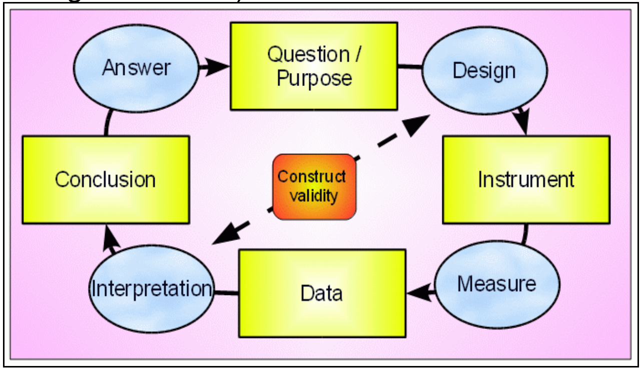


Figure 2: Assessment as a process

Any instrument, in fact any implementation of an idea, is a compromise between the ideal and the constraints of reality. Assessments in this sense are no different. They are an attempt to translate a desired measurement into an instrument. When this instrument is replaced with another, for instance in the transition to computer-based assessment, we should be careful with simply implementing the old instrument into the new medium. Doing so would mean compromising the instrument with both the limitations of the old medium, and those of the new. Instead we should look again at our ideal, the 'question' in figure 2, and see how it can be best implemented in the new medium.

Reconsidering how we can translate the original question into a suitable design becomes especially important when assessments haven't been designed to be authentic, but instead measure capability through a proxy, as is often the case. If test authenticity is considered, then the increasing ubiquity of technology alone (see Björnsson, 2008) certainly supports a transition to computer delivered tests. But technology might also provide us with different affordances. Examples of new types of assessment that can be supported by technology include interactions such as games and simulations, but also scoring based on response times, instead of solely on the responses themselves (Kyllonen, 2008). There is a lot of support for the idea that computer-based assessment will allow us to assess competencies that cannot be assessed in a paper and pencil test, for instance in Lee (2008) who states that "Computer-based assessment is expected to improve the assessment of scientific inquiry because it can allow students to make observations, manipulate variables, perform examinations, and gather data, tasks that are not possible with the paper and pencil test". Wilhelm (2008) however cautions us to not light-heartedly infer differences in what is measured, simply on the basis of cognitive labels that we assign to this measurement. The fact that we call something different, or perceive it as different, doesn't necessarily mean it actually measures something unrelated. The reverse is also true, and measurements that may look the same or are labelled to measure the same, may actually draw upon different constructs and competences.

Accuracy

Thompson & Weiss (2008) set out that the 2 main reasons to implement CAT are efficiency and effectiveness. Efficiency is of course discussed above, and much of the same considerations of the transition to CBA apply to the transition to CAT. Effectiveness however is different, and defined as a higher degree of precision of the measurement, and a more equal degree of precision for candidates of all abilities. There are 2 important caveats to be made around this improved precision.

Firstly, as a medium, *Computer Adaptive Testing* (CAT) presents us with a serious design challenge. The models used for item selection in a CAT require each question to behave in a very particular way, usually in accordance with the Item response theory (IRT). This requirement actually significantly limits the options that are available in the design of the instrument. Unfortunately, it is often the more authentic and complex question types that are excluded by this requirement. As such, the compromises that are made in the design phase can often be significant. The risk therefore exists that a CAT will only allow you to measure the wrong construct, albeit very accurately and reliably. More recent work in the use of polytomous items (Wandall, 2008) Might partially remove these limitations.

Secondly, it is important to realise that while measurements in themselves are never wrong, they also have little value. It is our understanding of what these measurements mean that is valuable, but also subject to question and prone to error (Wiliam, 2001). Improved accuracy in a measurement in itself therefore is not valuable, if it doesn't lead to an improved interpretation of these measurements. Or, to translate this to the model in figure 2, the capability of the assessor to interpret the results correctly, based on an understanding of the design of the assessment. If these inferences are crucial, the primary concern when designing an assessment should be that it delivers data to the person making these inferences in a format that allows this person to do so appropriately. With an instrument as complicated as a CAT, one has to ask if a valid interpretation of the measurements of an instrument as complex as a CAT be made by non-specialised people such as teachers, students and parents, or if this is an instrument that is suitable only in the hands of specialised professionals.

And there are other, arguably better, ways to improve accuracy. Shavelson, Baxter, & Pine, (1992) show that a reliable assessment of a learners competency requires a multitude of tasks, probably around 6 to 10 depending on the nature of the subject. The individual accuracy and reliability of these tasks might not even be that important. It is the spread over time, types of tasks and areas within the domain that give the aggregated results a comprehensive reliability and validity beyond what can be achieved with any single test.

Reconsidering assessment

The discussion above assumes that our current assessment system already is a measure of quality. But there are many signs that our assessment system is anything but satisfactory, and that in fact it might be broken. In the UK the inadequacy of the university degree system has been recognised for years, but tentative trails to provide some change and transparency have only commenced very recently (Shepherd, 2008). Whether employers want to wait for the sector to redeem itself remains to be seen. Some employers have already started their own education initiatives, which in the case of the 'McDonalds A-Levels' even led to the employer being certified to award accredited A-level diplomas.

With all these obvious signs of failure, or at least concern, around current practice, focussing on 'concurrency' might be the worst thing we could do. After all, we would simply be copying the mistakes and inadequacies of the existing system. Instead we should have a critical look at what it is that we should be learning and teaching, and what criteria to define for success. When students exchange information and ideas on their assignments via *Facebook*, should we be cracking down on this as plagiarism, or should we consider rewarding collaboration?

Conclusions

In designing an assessment system that is fit for purpose, we need to first carefully consider this purpose and its stakeholders:

- What decisions do we expect stakeholders to make, leading to our objective?
- How do we gather and present information in a way that it enables our stakeholders to make these decisions?
- What consequences might the act of gathering this information have?
- How do the consequences of the act of measuring compare to the benefits of the decision the measurement allows us to take?

Some general principles that will be helpful:

- Authenticity. The more authentic an assessment is, the more likely it is to be valid.
- Transparency. The more transparent an assessment process is, the more likely the outcomes are to be understood and accepted.
- Multiplicity: Every assessment method has its strength and weaknesses. The variety of assessment methods and moments is more important than their individual validity and reliability.

References

Björnsson, J. (September 2008): CBAS assessment and lessons learned with national representatives from Iceland. In The Transition to Computer-Based Assessment: Lessons Learned from the PISA 2006 Computer-Based Assessment of Science (CBAS) and Implications for Large-Scale Testing.

BBC (13 May 2008) Tests 'damaging' to school system. Available at: http://news.bbc.co.uk/1/hi/education/7396623.stm [Accessed November 24, 2008].

Bridgeman, B. (September 2008): Computer-Based Testing in the USA , in Scheuermann, F. & Björnsson, J. (Eds), 2008. The Transition to Computer-Based Assessment. Lessons learned from the PISA 2006 Computer-based Assessment of Science (CBAS) and implications for large scale , Luxembourg: Office for Official Publications of the European Communities.

Kyllonen, P.C. (September, 2008). New Constructs, Methods, & Directions for Computer-Based Assessment. , in Scheuermann, F. & Björnsson, J. (Eds), 2008. The Transition to Computer-Based Assessment. Lessons learned from the PISA 2006 Computer-based Assessment of Science (CBAS) and implications for large scale , Luxembourg: Office for Official Publications of the European Communities.

Mee-Kyeong Lee. (September 2008): Computer-based Assessment in Science (CBAS):
Experiences in Korea, in Scheuermann, F. & Björnsson, J. (Eds), 2008. The Transition to Computer-Based Assessment. Lessons learned from the PISA 2006 Computer-based Assessment of Science (CBAS) and implications for large scale , Luxembourg: Office for Official Publications of the European Communities.

Shavelson, R. J.; Baxter, G. P. & Pine, J. (1992). Performance assessments: political rhetoric and measurement reality. Educational Researcher, 21 (4), 22-27.
Shepherd, J., 2008. Testing times for degree as report card trial begins. Guardian. Available at: http://www.guardian.co.uk/education/2008/oct/21/report-cards-new-grades [Accessed November 28, 2008].

Møller Andersen, A.; Sørensen, H.(September 2008): How do Danish students solve the PISA CBAS items? Right and wrong answers in a gender perspective, in Scheuermann, F. & Björnsson, J. (Eds), 2008. The Transition to Computer-Based Assessment. Lessons learned from the PISA 2006 Computer-based Assessment of Science (CBAS) and implications for large scale , Luxembourg: Office for Official Publications of the European Communities.

Wandall, J. (September 2008): Tests in Denmark – CAT as a Pedgogic Tool, in Scheuermann, F. & Björnsson, J. (Eds), 2008. The Transition to Computer-Based Assessment. Lessons learned from the PISA 2006 Computer-based Assessment of Science (CBAS) and implications for large scale , Luxembourg: Office for Official Publications of the European Communities.

Wilhelm, O. (September 2008): Issues in Computerized Ability Measurement:
Getting out of the Jingle and Jangle Jungle, in Scheuermann, F. & Björnsson, J. (Eds), 2008. The Transition to Computer-Based Assessment. Lessons learned from the PISA 2006 Computer-based Assessment of Science (CBAS) and implications for large scale , Luxembourg: Office for Official Publications of the European Communities.

Wiliam, D. (2001) An overview of the relationship between assessment and the curriculum, in: D. E. Scott (Ed.) Curriculum and Assessment, pp. 165–182 (Westport, CT: Ablex Publishing).

The author:

René Meijer
University of Derby
Kedleston Road
Derby, S8 0RA, UK
e-mail: r.meijer@derby.ac.uk

René Meijer currently heads the Educational Development Unit at the University of Derby, which focuses on the development of electronic tools and content in support of learning, teaching and assessment. René is the technical lead and designer of the e-APEL project, which is developing a framework for the electronic support of the accreditation of prior experiential learning. Other interest include collaborative and peer assessment and the personalisation of learning. René is a fellow of the Higher Education Academy.

Transitioning to Computer-Based Assessments: A Question of Costs

Matthieu Farcot & Thibaud Latour
Centre de Recherche Public Henri Tudor

Abstract
Transitioning to Computer-based assessment (CBA) from paper-and-pencil (P&P) testing introduces strong differences in terms of costs. This article proposes a general framework dedicated to the analysis of costs as a support to decision-making. We illustrate the framework using the item production process and demonstrate that even at this early stage, CBA can offer sustainable cost advantages when compared to standard P&P approaches.

Transitioning from Paper-and-Pencil (P&P) testing to Computer-based Assessment (CBA) is a popular topic currently discussed among educational large-scale assessment and school monitoring communities. Besides the intensive debates about educational and psychometric issues, assessment specialists and policy-makers recurrently raise the same questions: What are the costs of transitioning from P&P testing to CBA? How could such costs be managed? And most importantly: does switching to CBA *really* reduce costs? Indeed, deciding to shift from P&P to CBA is not a trivial question.

The objective of this paper is to propose a general and simple decision-making framework which would allow comparing relative-cost elasticity of key factors and processes induced by paper-and-pencil and/or Computer-Based Assessment technologies. Based on various scaling variables, this decision-making framework considers each individual processes related to Computer-Based Assessment versus paper-and-pencil testing (such as item creation, test delivery, subject management, scoring…). This model is based on (the) hypothesis (of) several potential technological scenarios. It is, however, limited in its analysis specifically to the specific process of item production. This choice was made because of the essential nature of this process that is often considered as one of the most time consuming among assessment-related activities.

The authors demonstrate with this model that the cost elasticity of this process is not technology neutral, and even such early step in assessment procedures can potentially benefit from computing technologies to reduce induced costs. This model therefore demonstrates that any cost/benefit ratio is strongly dependant on scaling variables. Depending on the amount and complexity of items, taken as scaling variables, the most cost efficient technology might change.

Opposing the costs of CBA & P&P

The issue of measuring the cost of introducing CBA has hardly ever been addressed in the (industry) literature. Indeed, most often, the transition to CBA is justified by educational arguments and rarely uses purely economic justifications (Ricketts et al., 2003). Yet, most authors claim – in addition to other direct benefits related to education and assessment – that it can enable significant cost reductions, mainly due to positive externality effects and scale-based savings. On the contrary, other authors have stressed the fact that switching to CBA can represent a costly operation. For example, the first generation of computerised assessment, which consisted of rough transpositions of P&P tests into computerized counterparts, has been reported to increase costs in high stakes assessment (Bennett, 1997). Even if the next generation-computerised assessments have been reported to be less expensive, it remained unquestioned that CBA was still be considered more costly than P&P (Jamieson, 2005).

The cost of tailor-made test development has also been highlighted in the framework of large-scale studies (Lennon et al., 2003; Murray et al., 2005). This higher cost of computer-based tests with respect to P&P testing has been reported as a potential obstacle to the development of computer-based assessment in schools (Bennett, 2001). This case occurs especially in “one-shot” scenarios where no learning effects could be gathered. Among the possible origins of such costs, the commercial software and the related cost of the licence constitute one of the major identified economic barriers for the adoption of CBA (Conole and Warburton, 2005).

Obviously, cost considerations play an important role in decision-making relative to the deployment of computerised assessments. In a paper specific to the impact of CBA on Higher Education Institutions, the importance of constructing methods for evaluating the cost of CBA has been stressed (Bull, 1999), but this analysis is context-sensitive.

As exposed earlier, few studies have been published in the past involving clear and reliable empirical data on this specific issue. Among the few available studies, a cost estimate comparison between two different technologies enabling the scoring of tests based on Concept Maps can be found in a report from the University of California (O'Neil and Klein, 1997). Later, a cost benefit evaluation of the introduction of CBA into a mathematic course has demonstrated significant reduction of time spent by staff on preparing and scoring the examinations. The authors also concluded that CBA provides significant cost cutting opportunities (Pollock, Whittington, Doughty, 2000). Relative to this specific question of cost management in the debate opposing CBA to P&P, Ricketts et al. (2003) proposed a framework to evaluate costs and benefits. They provided lists of activities that might lead to detailed cost calculation. Unfortunately, the paper addressed only the surface of the problem, and while bringing valuable hints to initiate the debate; they did not provide a complete abstract and transposable framework.

Identifying the Target in Computer-based Assessment Diversity

Assessment in general and computer-based assessment in particular takes place in heterogeneous contexts and situations. Intrinsically the world of assessment and CBA bears a large and potentially intricate space of variability due to this heterogeneity of induced actions. The transition from P&P to CBA must take into consideration the different dimensions of such diversity, and the dynamics to which they relate to, such as network effects and learning curves. The decision rationale between a series of potential scenarios depends strongly on where the actual assessment context fits in this space: as such, one case can hardly be compared to another.

The authors of this article have identified the diversity of possible assessment situations under three dimensions: the context, the content, and the container.

The context

Very schematically, assessment in general and, more specifically, CBA can serve diverse purposes. Assessment can either be used as a tool dedicated to measure a current situation (as for summative assessment), or to anticipate and drive future evolution (as for formative assessment). This purpose can be applied at different holistic levels, from individual to system-wise. At the system level, the purpose may be to survey the performance of a population using a representative sample (PISA survey), or to monitor the expected evolution of the system, potentially by evaluating the full population (school monitoring).

The type of assessment needs will also strongly impact the decisions one will make when selecting the appropriate assessment implementation scenario. The space of needs encompasses the steering level ranging from the single individual to the entire system being considered, the chronological dimension (depicting the individual or the system evolution timeline) and finally the competency dimension (represented by the nature of the evaluated skill or competence).

Other factors directly impact the context of use and are related to the market targeted by the assessment. At the macro-level, such segments can be nation-wide educational programs in general or global socio-economic analysis. At a meso and/or micro level, such segments can be research in psychology, social and educational sciences, human resource business, regulation and/or legal certification…

Any assessment procedure will have to take such intrinsic particularities into account when designing assessment instruments and processes. Such particularities impact the cost function of each alternative technology.

Finally, there is the important issue of differentiation among contexts as they relate to the spatial location of the assessment, which impacts the resource needs (both internal and external), and therefore influences strongly the cost-based evaluation of CBA versus P&P. As an example, from schools, system-wide to more limited classroom-wide, the model will naturally be completely different and the conclusions will need to take this into account.

The content
Independently from the context of testing, one of the main factors impacting the assessment of costs and induced benefits relates to the measurement instruments. The choice of the testing sequence algorithm has a significant impact on cost elasticity. For example, Complex conditional branching and adaptive testing requires an increased number of items. Also, the type of items considered, the nature of the interaction (multiple choices, open-ended questions, …), and the means of scoring are critical factors that need to be taken into account when comparing the costs and benefits of heterogeneous technologies such as CBA or P&P.

Relative to the same issue, depending on the context and objective of the test, different kinds of reports might be produced, from very complex statistical analysis and figure generation to simple charts and histograms, which also naturally impact the content.

This brief summary underlines the heterogeneity of the situations related to technology-induced costs with respect to the content, *i.e.*, the instrument as such.

The container
Although CBA clearly relies on software developments, the software needs are heterogeneous since their architecture and their relative delivery options range from laptop-based to networked applications. This heterogeneity is also impacted by the use of the developed tests. In the case of a single purpose test, the later would not change over time, and a hard coded application would probably be among the most efficient solution. Extendable generic and fully interoperable applications on the other hand, fit particularly well among situations where the testing instrument undergoes disruptive changes over time.

These elements impact the reusability factor of assessments in a strong manner, which in turn impacts the cost.

Another important issue for the container concerns the software's licensing model. The type of licensing scheme applied to the software application will naturally impact the cost-based perspective. An Open Source licensing scheme applied to a CBA platform will grant developers the freedom to generate a clear cost advantage versus a proprietary software model (Farcot, Latour, 2008).

Instantiating the model using our Return from Experience

This section, based on Martin and Latour (2006), illustrates the reasoning underneath the hypothesis upon which our model is based and will be exposed in the next section.

Large-scale Assessment of Scholastic Aptitudes in School Monitoring Programmes: the Luxembourg case (2006-2009)
The Luxembourg State is currently implementing school-monitoring programme based on CBA. In order to be able to deliver a large number of tests under restricted time frames, a dedicated hardware infrastructure has been pre-tested in July 2006 and a first countrywide measurement campaign was then made in October 2006. A second successful countrywide campaign was done in July 2007, and this program will now be followed on a yearly-basis

The hypothesis that we shall use for our model took into account the following context: pre-testing of a language literacy evaluation delivered in-line in the form of C-Test items to 4000 students in several schools over the country. The test sessions were performed using the existing IT infrastructure of the schools. The field-trial campaign took place over 8 days. The daily schedule consisted of 4 synchronised sessions. During this period an average of 400 tests were executed per day with a peak of more than 1000 simultaneous test executions.

PISA 2009 international survey (Electronic Reading Assessment)
The TAO platform, an Open Source CBA solution created by the CRP Henri Tudor and the University of Luxembourg, is currently being used to implement the instruments of the optional Electronic text Reading Assessment (ERA) to be delivered on a computer platform in the framework of the PISA 2009 international survey.

To achieve the task, a specific stimulus emulating a web browser and a mail client was developed by a third party software service company under the supervision of the DIPF (*Deutsches Institut fur Internationale Pädagogische Forschung*). The stimulus enables the recording of all testee actions during the test execution. This new stimulus has been developed together with a dedicated authoring application. In addition, several client-side plug-ins have been provided to fine tune the test management and appearance.

On the server side, a simplified web-based authoring item has been proposed to the developers of the participating countries. In addition, a series of extensions supporting the complex translation process of tests and items are currently under development.

Finally, as the PISA 2009 context demands that the test must be executed on an existing school infrastructure, a CD solution for test delivery has been implemented.

Luxembourg Ministry of Education and Professional Training: Mathematic assessment

The national Ministry of Education of Luxembourg used a CBA platform in the framework of mathematical assessment of Luxembourgish pupils. This computer-aided test used P&P instruments and a CBA platform to manage the process. Dedicated software components were developed to enable teachers to evaluate a students' answer for each item of the test and to consolidate data and perform statistical analysis of the results at classroom, school, and national levels. Background batch processes were used to produce reports to be sent to the different stakeholders.

Learning Tool in Mathematics Classes (2006-2007)

This use case (called CAMPUS for *Computer-Assisted Mathematical Problem Understanding and Solving*) has been primarily conceived as a formative assessment tool. It provides a structured environment that assisted the learner in the process of mathematical problem solving. The tool provided the user with a highly interactive environment in order to represent graphically mathematical operations. This t environment was specifically developed for this purpose.

Dynamic Evaluation of Scientific Literacy (2006-2007)

This use case (called CASCADE for *Computer-Assisted Scientific Competencies Assessment and Dynamic Evaluation*) demonstrates the added value of computerized assessments, especially for a more process-oriented and dynamic approach.

It consisted of a two-phase testing procedure whereby the current knowledge state of the testee is evaluated through a series of multiple-choice questions, and then reviewed by the testee using multimedia based sources of information. This instrument was, as in the previously exposed case, specifically developed for this purpose.

Adaptive Placement Test for the Assessment of Language Skills (2005–2007)

The *Centre de Langue Luxembourg* (CLL) is a major language training school in Luxembourg, which provides training in a large number of languages. Traditionally, students are assigned to a correct group level through a paper-and-pencil test dedicated to listening and reading proficiency, followed by a personal interview.

The CLL experimented with adaptive placement tests for German, French and English languages. These tests were adaptive and consisted of item banks of about 130 time-limited calibrated items per language ranging on the European reference framework scale (A1, A2, B1, B2) according to the 2-parameters of the Birnbaum IRT model.

During the first deployment phase, the German test was made available to the CLL computer pool. Teachers at the CLL then developed by themselves the items for French and English tests. These items were calibrated by the University of Luxembourg using results obtained from paper-and-pencil test sessions.

Using CBA has considerably reduced the organisational complexity of paper-and-pencil tests as well as the test duration, specifically due to the adaptive testing.

This use case allows us to compare P&P and full CBA situations.

The Model

Our model is based on a theoretical analysis of the previously exposed use cases and for which a qualitative evaluation of several key factors characterising the cost functions used in P&P or CBA has been made.

For the sake of clarity, we have organized the cost-structure of our model by regrouping three main categories, which are logistic, opportunity and organizational costs.

As illustrated in the figure below, Logistic costs can either be internal or external, as they would either relate to the internal infrastructure needs or externally-produced consumables.

Opportunity costs relate to specific risk-management issues. Finally, we incorporate organizational costs that relate to the workforce needs.

Since other kind of costs might impact each of the proposed categories, we have listed miscellaneous costs independently for each category.

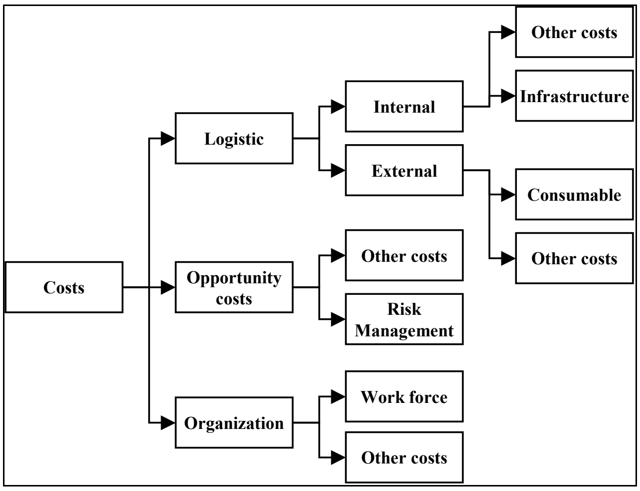


Figure 1: Model cost structure

As exposed earlier, our model illustrates one specific scenario related to items production.

The implementation of the item production process can be characterized with respect to various technological scenarios.

Paper-and-pencil (PP): In this scenario, the global assessment process is based on paper procedures and instruments. Computers and IT are possibly used to support the process. However, IT tools in this scenario are not dedicated to the assessment business, neither to support back-end, management, and delivery procedures, or as part of an assessment instrument. According to each particular process, many variants can be observed. For instance, scoring can manually be made and encoded in a database, or coded automatically with optical devices and directly encoded in the database. Even if some technology supports the process here and there, the overall scenario relies on paper-based processes.
For the specific scenario and related processes of item production, this technology is illustrated by the P&P placement test for the assessment of language skills.

Computer-aided (CA): This scenario is very similar to the P&P scenario. However, the items are managed in a centralised system manipulated by the authors themselves (contrary to P&P scenario where IT manipulations are made by dedicated IT people). The items therefore are managed centrally in electronic form (PDF and meta data for instance), but are still produced on paper for discussion and usage.
This scenario relates to our use case concerning the Luxembourg ministry of education and the professional training mathematic assessment.

Computer-based with taylor-made system *(TM):* In this scenario, the items are in an executable electronic form (this means that they are executed on a computer when the subject interacts with them). In their final form they are not directly produced by authors. On the contrary, they are produced by authors on office tools (as mock-ups or pseudo-specifications), and are re-programmed in their final form by IT specialists. The management of the resulting items in electronic final form is made in a central repository, as in the CAT scenario. There is no real template notion since final items are pieces of tailor-made programs created almost from scratch.
This scenario is illustrated in part with the PISA 2009 ERA example, the learning tool in mathematic classes, and the dynamic evaluation of scientific literacy.

Computer-based with general platform (PF): In this scenario, the items are produced by the authors using a dedicated authoring tool enabling them to create the items directly in their final executable form. The overall management of items is directly made within the platform; there is no circulation of items through mails, or other means. The system requires intensive training of the authors, as well as initial developments to set up the required framework. However, most item-functions have been isolated in the platform and can be reused as is. This is particularly true when there exist a few well-defined templates.
This last scenario is illustrated by the large-scale of scholastic aptitudes in Luxembourg school monitoring programs, the adaptive placement test for the assessment of language skills, and to a certain extent PISA 2009 ERA

Instantiating the model – workforce factors for item production

We shall in this section illustrate the costs related to the workforce needs. The general shape of the cost structures related to the amount and complexity of items is shown below in figure 2.

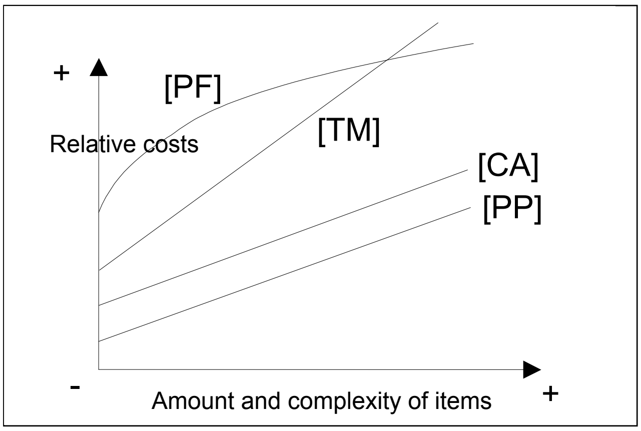


Figure 2: Relative cost evolution of workforce factors with respect to the amount and complexity degree of items

PP scenario: The training needs of authors related to item design create a first initial set of fixed costs, which are then impacted by a constant marginal price per item increase (basically related to the time spent on it). Such marginal costs include the item review process.

CA scenario*:* Authoring and designing the items creates a first initial fix cost, evolving from then on to a constant marginal price per item (basically related to the time spent on it). As in previous case, the review is included in the marginal cost. We assumed that the training cost of CA is probably higher than PP due to the management of extra IT system needs. Contrary to PP scenario, specific IT tools induce a learning curve and associated costs.

TM scenario: After an initial need for training, the costs are then raised under a constant marginal cost (equivalent to PP or CA) related to item design needs. In addition, there is an extra marginal cost for the final programming. A third marginal cost-impacting factor arises from the iterative loop between the developer and the programmer to create the items. Some modification may be necessary and a new cycle may be triggered, generating new workforce related costs as a retroactive process.

PF scenario: There is a strong training need because of the special authoring tool and because of the fact that the authors will also need to manage a large deal of the process. However, since templates and reusable libraries can be used, the marginal cost decreases as the amount of items grows. There is no modification cycle between authors and developers since the author can directly observe what he has produced. There is also a dynamic effect induced to learning curves, as the author is getting acquainted to the system.

Instantiating the model – Organisational related factors for item production

Organisational costs include workforce and other endogenously-related costs enabling the item creation. The latter costs are deeply impacted by the nature of the technology used, as discussed hereafter and illustrated in figure 3.

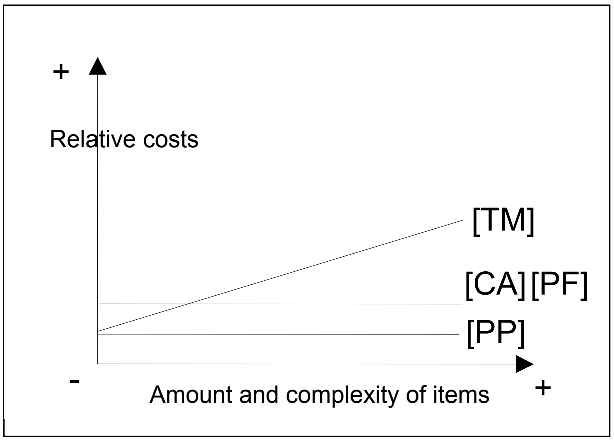


Figure 3: Relative cost evolution of Organizational factors with respect to the amount and complexity degree of items

PP scenario: Among the PP scenario, the organisational costs mainly consist of the overall management of the item production team. It does not depend on the number of items, but on the complexity of the process, and therefore is a fixed parameter.

CA scenario: The organisational costs mainly consist in the overall management of the item production team. As for PP, it does not depend on the number of item, but on the complexity of the process. The technology does not affect this cost function, and is equal to the organisational costs of the PP scenario. However, we introduced an extra fixed cost covering the management of the IT administration team.

TM scenario: The organisational costs are mostly composed of overall management costs due to the item production team. As with the previous two scenarios, it does not directly depend on the number of items, but on the complexity of the process. There is however an extra cost introduced by the outsourcing management of the item diffusion via media. This extra cost is recurrent and depends slightly on the number of item. This yields an additional constant marginal cost.

PF scenario: The organisational costs mainly occur in the overall management of the item production team. As for all previous scenarios, it does not depend on the number of items, but on the complexity of the process. The initial cost should include the management of the administration team as with the CA scenario. It is therefore perceived as slightly higher than PP related cost function.

Instantiating the model – Opportunity related factors for item production

Opportunity costs consist of risk and related insurance management expenditures.
The general shape of these cost structures as related to the amount and complexity of items is illustrated in figure 4.

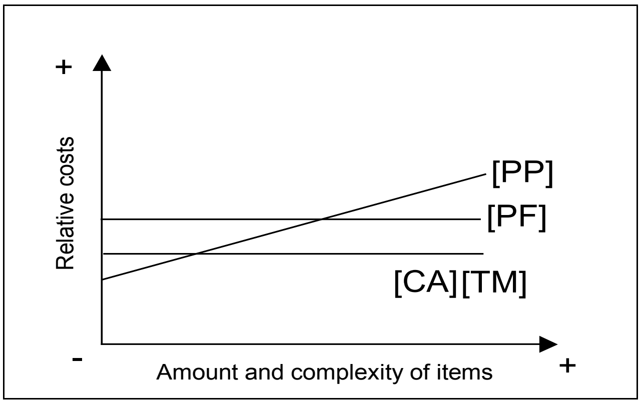


Figure 4: Relative cost evolution of opportunity factors with respect to the amount and complexity degree of items

PP scenario: The management of risks and insurance policies consist in protecting the items created against divulgation, degradation, and loss. It should in addition prevent action that would disable permanently or temporarily the use of the items. This represents an overall cost structure that should have a large initial fixed cost contribution, which we then assume as endogenous to the amount and complexity of items due to the physical manipulation of items (*i.e.*, insurance associated to logistics).

CA scenario: The cost structure is essentially based on the protection of the central repository where items are stored in an electronic format. The exchange of items is made electronically and the marginal cost per item can therefore be considered as nil. Differently, the cost of protecting the IT infrastructure might be slightly higher due to additional technical constraints (the physical protection of the infrastructure is similar to PP, plus an IT security contribution added to the cost). Hence, cost function can be treated as a constant with no marginal cost impacts, but with initial costs higher than for the PP scenario.

TM scenario: The cost of risk management is similar to the cost-related hypothesis of the CA scenario.

PF scenario: In this scenario, the cost of risk management is slightly different from the three other scenarios. Since many users access the platform in interactive sessions, the risk related to IT security is increased when producing the items. This is particularly true when multi-site access is provided, possibly through the web. If we can assume a rather centralised access to IT infrastructure in the CA and TM scenarios, this assumption does not hold in PF scenario. Hence, there is an extra fixed cost for wide access to IT infrastructure and related security needs. This cost does not however depend on the number of items, but on the location and amount of users, therefore inducing no marginal extra costs. By hypothesis, we assimilated the opportunity cost function for the PF scenario as constant and higher than CA and TM.

Instantiating the model – Infrastructure related factors for item production

We shall now discuss the cost function for each of the scenarios related to the infrastructure costs. We include among infrastructure expenditures all hardware (servers, laptops…) and real-estate (rents of offices…) related costs. Consumable costs are excluded.

PP scenario: The minimum infrastructure costs relate to housing, hardware and other basic organisation infrastructure needs. We have considered the latter as a fixed initial cost. Infrastructure-related costs vary with respect to the amount of employees and are not perceived as correlated to the number or complexity of items.

CA scenario: The minimum infrastructure costs are the same as for the PP scenario. It is considered as a fixed initial cost. This cost also includes a special infrastructure enabling the central management of items. This additional IT infrastructure makes the initial cost higher than for the case of PP. As stated previously, this cost varies as a function of the number of employees and does not depend on the number or complexity of items.

TM scenario: As for each of the two previous scenarios, infrastructure-related cost functions remain fixed. We estimate by hypothesis that such costs are about the same than previously.

PF scenario: Finally, as discussed above, the minimum infrastructure costs remain the same. However, to this initial fixed and technology-neutral cost, one must add the specific cost of the platform and its deployment, maintenance, and exploitation. This cost should also include the training of administrators (not included in the workforce cost which is allocated to authors producing the items). As such, the infrastructure cost has a high initial fixed cost, higher than PP, CA, and TM. This cost then varies with respect to the number of involved employees and does not depend on the number of items.

Instantiating the model – Consumable related factors for item production

The general shape of the consumable-related cost functions with respect to the amount and complexity of items will now be discussed below.

PP scenario: Related to the PP scenario, no assessment-specific consumables need to be considered by hypothesis. The latter are represented by an overall fixed cost. In addition, since everything is produced on paper, including all activities pertaining to the production of items (reviewing...), we estimated a significant marginal cost increase depending on the number and complexity of items.

CA scenario: This scenario is very similar to the PP scenario regarding consumable costs for item production. However, the central IT management should decrease slightly the level of consumable-related fixed costs when compared to PP.

TM scenario: Since all the design related to test and assessment item production is very similar to PP, we can assume that the consumable cost structure is approximately the same as for PP.

PF scenario: Finally, in this scenario, all item production is expected to be computer-based. We can assume extremely weak consumable cost. We assume no marginal costs for this scenario, since the increase of the amount of items can be done on a specific computer. Above all, PF cost function is characterized by the weakest fixed consumable-related costs of all scenarios.

Exploring the Model

Based on our initial assumptions, the scenario related to PP induces the lowest initial fixed costs for item production. This scenario remains highly competitive for a low amount and complexity of items. However, this scenario will end up being the most expensive as the amount and complexity of items increases.

TM and CA scenarios each include initial fixed costs, which are higher than PP. However, due to a lower item complexity elasticity, they remain cost-competitive and will end up being cheaper than PP for medium to high values of the scaling variable.

Finally, the PF scenario is impacted by an increasing marginal return, *i.e.*, should initial costs expected to be the highest of all, the PF cost function will induce the lowest costs for high level complexity. As a conclusion, the PF scenario is the most cost effective scenario for high number and item complexity.

As we can see, comparing cost structures from computer-based assessment to paper-and-pencil is not a trivial task. The hypothesis induced among each scenario restrained us from giving a context independent and definitive conclusion.

However, even among the first step of test and assessment creation, namely the item development, we can see that the technological choices should strongly impact the cost structure.

Conclusion

Usually, two types of arguments are put forward for the benefit of CBA versus P&P. Firstly CBA improves time-to-delivery. Secondly CBA reduces mid- and long-term costs (by generating economies of scale). However, as illustrated through the 3 CBA-related scenarios that have been developed in this model (CA, TM and PF), CBA is multiform. As a matter of fact, the diversity of CBA is so large that searching for a unique, general and transposable answer concerning cost efficiency of CBA as opposed to P&P is misleading. On the contrary, deciding between different CBA scenarios and P&P scenarios should be scrutinized on a case-by-case basis. The framework hereby proposed can support a structured decision related to this issue.

This model only illustrates one specific assessment process, the first initial instantiation linked to item creation. The Cost functions illustrated should be completed by those of all other assessment processes such as producing the tests, managing the subject and related groups, delivering and executing the tests, and finally analysing and reporting the results.
A thorough modelling of all processes taking into account all related scaling variables would be required to obtain a more exhaustive picture.

Acknowledgement

The authors would like to thank Donald Martin (CRP Henri Tudor) and Friedrich Scheuermann (JRC, EUC) for their help, advices and support.

References

Bennett, R. (1997) Speculations on the Future of Large-Scale Educational Testing, ETS Research Report RR-97-14, Educational Testing Services, Princeton, NJ, USA
Bennett, R. (2001) How the Internet will help large-scale assessment reinvent itself. Education Policy Analysis Archives, vol 9, p. 5
Bull, J. (1999) Computer-Based Assessment: Impact on Higher Education Institutions, Educational Technology & Society, vol. 2, Nr. 3
Farcot, M. and Latour Th. (2008) Open Source and Large Scale Computer Based Assessment Platform : A real Winner. In Scheuermann, F. and Guimaraes Pereira, A. (eds.), Towards a Research Agenda on Computer Based Assessment, Challenges and needs for European Educationnal Measurement, JRC Scientific and Technical Reports, EUR 23306 EN
Conole, G. and Warburton, B. (2005) A review of computer-assisted assessment, ALT-J, Research in Learning Technology, Vol. 13, No. 1, March 2005, pp. 17–31
Harold F. O'Neil Jr., Davina C.D. Klein (1997) Feasibility of Machine Scoring of Concept Maps, CSE technical Report 460, Center for the Study of Evaluation, University of California, USA
Jamieson, J. (2005) Trends in Computer-Based Second Language Assessment. Annual Review of Applied Linguistics, vol. 25, pp. 228-242
Lennon, M., Kirsch, I., Von Davier, M., Wagner, M., Yamamoto, K. (2003) Feasibility study for the PISA ICT litteracy assessment, Report to Network A, OECD
Martin, R., Latour, Th., Burton, R., Busana, G., Vandenabeele, L. (2006) TAO: Several Use Cases of a Collaborative, Internet-Based Computer-Assisted Testing Platform, in Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2006
Murray, T.S., Clermont, Y., Binkley, M. (Eds). (2005) Mesurer la litératie et les compétences des adultes : des nouveaux cadres d'évaluation. Statistics Canada, Ottawa, ONT
Pollock M.J., Whittington C.D. and Doughty G.F. (2000). Evaluating the Costs and Benefits of changing to CAA. Proceedings of the 4th CAA Conference, Loughborough: Loughborough University
Ricketts, C.; Filmore, P., Lowry, R., Wilks, S. (2003) How should we measure the costs of computer aided assessment? in: Proceedings of the 7th CAA Conference, Loughborough: Loughborough University

The authors:
Matthieu Farcot, Thibaud Latour
CRP Henri Tudor
29, Avenue John F. Kennedy
L 1855 - Luxembourg
E-mail: matthieu.farcot@tudor.lu
Thibaud.latour@tudor.lu
WWW: www.tudor.lu

Matthieu Farcot holds a PhD in economics, defended in 2006 at the Bureau d'Economie Théorique et Appliquée (Beta) of the Université Louis Pasteur, Strasbourg, France. This PhD focused on the topic of Intellectual Property Dynamics within the Software Industry. He holds Masters in economics, management, and law. In particular, he holds a master in Law (LLM degree) focused on European Law and Intellectual Property obtained in 2008 from the Centre d'Etudes Internationales en Propriété Intellectuelle (CEIPI) of the Université Robert Schuman, Strasbourg. He also holds a double master in innovation economics and knowledge management (obtained from the Beta in 2002). He finally holds since 2005 a master equivalent degree in Intellectual Property Management obtained from the IEEPI (Institut Européen Entreprise et Propriété Intellectuelle), Strasbourg. Currently working within the valorization of ICT innovation Unit of the CRP Henri Tudor as a Project Manager, his field of study focus on intellectual property and legal issues related to IT based innovation and software developments. He also works on IT licensing and business models applied to free and open source projects, such as the design of a decision-support tool dedicated at choosing an Open Source license based on economic and management related considerations.

Thibaud Latour is the Head of the "Reference Systems for Certification and Modelling" Unit at the Centre for IT Innovation department of the Public Research Centre Henri Tudor. He is also Program Manager of a project portfolio dedicated to Technology-Based Assessment of skills and competencies. He obtained his M.Sc. in Chemistry in 1993 from the Computer Chemical-Physics Group of the Facultés Universitaires Notre-Dame de la Paix (FUNDP) in Namur (Belgium), working on conceptual imagery applied to supramolecular systems and in collaboration with the Queen's University at Kingston (Ontario, Canada). From 1993 to 2000, he participated in several projects exploring Artificial Neural Network (ANN) and Genetic Algorithm (GA) techniques and developing ad hoc simulation methods for solving complex problems. During the same period, he supervised a number of M.Sc. thesis work in Computational Chemistry. In 2000, he joined the Centre de Recherche Public Henri Tudor where he was in charge of an internal Knowledge Management project. In that context, he designed a knowledge base for collaborative elicitation and exploitation of research project knowledge. He is in now involved in several projects in the fields of Computer-Based Assessment, e-Learning, and Knowledge Management where Semantic Web, Knowledge Technologies, and Computational Intelligence are intensively applied. Thibaud Latour has also served in programme committees of several conferences such as ISWC, ESWC, and I-ESA and as workshop co-chair of the CAiSE'06 conference.

Shifting from paper-and-pencil to computer-based testing: Requisites, challenges, and consequences for testing outcomes
A Croatian perspective

Vesna Buško
University of Zagreb, Croatia

Abstract:
The paper focuses on prospects of moving from paper-and-pencil to computer-based testing, pointing also at basic conditions required to facilitate decision-making and bring about desired changes in the assessment, and consequently in teaching practices and essential learning and educational outcomes. Prevailing practices in psychological assessment and testing in Croatia are shortly outlined. Potential barriers in implementing ICT in the assessment processes are discussed including knowledge or informational, organizational, policy and financial issues. The reflections on benefits and likely obstacles to transition to computer-based assessment are exemplified by recent experiences from large-scale testing projects, such as national curriculum tests and PISA 2006 in Croatia.

Prospective benefits of the use of information and communications technologies (ICT) in psychological and educational assessment practices are undoubted and obviously numerous. Many of these have repeatedly been stated, including gains in terms of efficiency of test administration and scoring process, accuracy in data coding, advances in precision of measurement, accessibility of additional information such as response times or process data, use of more diverse, richer and more attractive stimulus materials, and a range of dynamic and interactive items or tasks (e.g., Martin, 2008; Mead & Drasgow, 1993; Ripley, 2008; Wilhelm & Schroeders, 2008). Furthermore, apparent savings of costs related to test administration procedures, data entry, databases manipulations and analyses, as well as the issues of test security and impact on students' motivation, should also be put on the list of advantages of computer-based over the traditional paper-and-pencil modes of assessments (e.g., Björnsson, 2008; Pitcher, Goldfinch & Beevers, 2002; Yeh, 2006).

In view of these and other potential benefits of computer-based testing (CBT), along with ever-increasing technological advancements and an overall increase in ICT literacy skills, the change in testing mode seems to be inevitable. It may be worth noting that this venture, whenever admitted, is relevant to the extent that it contributes to the major objectives and purposes of the assessment.

Having these facts in mind, the present paper aims to reflect on some prerequisites for and likely obstacles to transition to computer-based assessment. The discussion will be exemplified by the experiences from Croatian educational context. Amongst the manifold important issues interconnected with these processes of transition in the assessment practices, the present paper will try to offer reflections and arguments which emphasize some critical requisites in this context as seen from the perspective of actual educational system in Croatia, but which can probably well apply to other countries with a similar level of socio-economic development.

There are several important prerequisites which should preferably be met prior to considering the prospects for alteration in the actual assessment practices, and certainly prior to making decisions on the mode and approach to assessment to be implemented. These prerequisites or challenges can be viewed and discussed depending on whether they are mainly associated with (1) examinees or students; (2) staff engaged in the assessment process, e.g., administrators in the testing procedures, teachers, school authorities; (3) research data, that is, available empirical evidence on test-scores equivalence for a given population; (4) other stakeholders, such as governing bodies, state and/or ministry officials, managers in business companies, project managers, etc.; and/or (5) other resources. Various sorts of complexity might appear when trying to address related requirements, primarily of methodological and technological nature. Possible obstacles in the enterprise of shifting from paper-based to computer-based testing have to do with all the stakeholders involved in this process, and can principally refer to knowledge or informational, organizational, policy and/or financial issues.

When speaking about challenges mainly pertaining to test takers, students or learners, potential barriers can be illustrated by different empirical data. For instance, according to self-report data obtained on the sample of Croatian participants (N=5242) in the last PISA cycle, large majority of students (92%) used computers, with 62.5% of the sample reported on using computers at home daily. Still, a considerable number of students used computers rather rarely, within a range of once a month or less to once or twice a week (Braš Roth, Gregurović, Markočić Dekanić & Markuš, 2008). Today, two years after the PISA study, these records would certainly look differently. However, the observed interindividual differences clearly raise the issue of fairness in case of computer-based assessment.

In general, these and other empirical data and professional experiences show that there are still sizeable individual differences in ICT skills among students, which can be noticed at different educational levels and in different age groups. Apparently, the differences are at least partly explainable by variations in socio-economic background of pupils. The differences can be found between certain types of schools, e.g., grammar vs. professional or other types of high-schools, between students living in urban and rural areas, between different regions within Croatia, and the like. The differences are logically expected to be more pronounced when between-countries comparisons are considered.

The issue of familiarity with ICT appears to be at least equally salient when considering the role of teachers and other academic staff engaged in the assessment process. To implement CBT procedures, assist students in taking tests, take advantage of the results and give feedback to learners, teachers and administrators in the assessment need to be sufficiently familiar with usage of ICT. Although many teachers use ICT regularly in their work, curricular and teaching activities, this is still not the case for many others, particularly when it comes to older teachers. Teaching staff with low ICT skills is less likely to actively use ICT in their ordinary classroom activities, and will also be less prone to participate in CBT or to assist in computer-based assessment surveys. They can hardly be expected to be enthusiastic about putting considerable amount of extra effort in conducting both paper-and-pencil and computer-based testing procedures. In addition, administering computer-based tests is usually connected with a range of specific organizational requirements, including e.g. various adjustments in teaching schedules, available room and personnel arrangements, which in contemporary settings of many Croatian schools is sometimes not really easy to accomplish.

The latter issue is further related to the availability of resources on the whole. Having appropriate space, time and skilled administration team at disposal has to do with the overall requirements for standardized testing conditions. These include adequate technical support, as well. As shown by the same PISA 2006 data (Braš Roth et al., 2008), 46% of Croatian schools at the time of the study were faced with a problem of insufficient number of computers in classrooms. School authorities of 28% of participating schools reported lacking or poor internet connections, while problems with inadequate or lacking educational software were reported for 64% of the schools. Again, the situation with ICT facilities in Croatian schools is expectedly better at the present. Nevertheless, it still appears to be far from reaching uniformity regarding computer hardware types and component performances, as well as software types and versions, both of which have been proven to be relevant features from the perspective of validity, fairness and standardization of assessment procedure in case of CBT (e.g., Bridgeman, 2008; Martin, 2008; Wang, Jiao, Young, Brooks & Olson, 2008).

There is a separate set of questions concerning the issue of comparability of test scores across different administration modes. As it can be derived from an extensive body of empirical data accumulated thus far on test mode effects, the answer to the issue of equivalence across test media is not simple. The degree of equivalence was shown to be dependent on test speediness, construct measured by a test, content domain level, technical aspects of item presentations, and other factors. Moreover, the determinants of the observed differences in scores obtained on two test forms appear to be test- and population specific, and also varying across software and hardware performances (Björnsson, 2008; Kim & Huynh, 2008; Mead & Drasgow, 1993; Wang et al., 2008; Wilhelm, & Schroeders, 2008). These findings clearly highlight the need for a thorough study of test media equivalence for any given population planned to be included in a certain computer-based testing program.

Obviously, all the aforementioned necessities require sufficient financial resources. Despite the certain costs savings associated with the implementation of ICT in the assessment practice, there are substantial financial means required for initiating system of electronic testing, the maintenance of computer infrastructure, research expenditure and the like.

Finally, a note on the relevant stakeholders' engaged in or responsible for decision-making in the area of psychological and educational testing policy should be offered. These stakeholders include educational authorities, governing bodies, state and/or ministry officials, managers in business companies, educational project managers, etc. Apart from different methodological, technical or technological aspects of particular assessment practices previously described in this paper, it seems important to remind to the role that policymakers and other related authorities play or could play in this enterprise. The discussion about and any action towards transition in the assessment mode practices make sense to the extent that the results of these endeavours, regardless of the mode favoured or applied, have an impact on the existing educational practice, and can add to its purposes and outcomes. Government and educational authorities can considerably contribute to this process by their full recognition of the relevance and implications of assessment, including the advantages of computer-based testing; by their interest and focus on the testing results, and adjusting their actions to match the results; by their readiness to initiate changes based on the assessment outcomes (for instance, by investment in courses for teachers on alternative teaching methods, by initiating educational reforms upon the insight into the testing results, by including ICT skills into school curriculum as required subject, and the like); and ultimately, by their willingness to invest into related projects.

The assessment practices pertaining to most research and applied areas of psychology in Croatia today predominantly include paper-and-pencil method of test administration. Aside from research context, the shift to computer-based and computer adaptive testing can apparently be expected soon within the field of professional selection and human resource management. An example is an online approach to vocational guidance with e-assessment of professional interests that has been applied for several years.

As far as the educational context is concerned, and large-scale testing programs in particular, the transition to computer-based testing does not seem yet to be a realistic goal. The present paper aimed to outline a range of requirements and challenges related to this major change in the testing mode practices. Different sources of difficulties are emphasized in the text including financial, policy, organizational, informational and fairness issues. Each of these issues should be acknowledged and addressed by the relevant stakeholders in order to make the process of shift to computer-based educational assessment a reasonable venture.

References

Björnsson, J. K. (2008) Changing Icelandic national testing from traditional paper and pencil to computer-based assessment: Some background, challenges and problems to overcome, in F. Scheuermann & A. Guimaraes Pereira (Eds) Towards a research agenda in computer-based assessment: Challenges and needs for European Educational Measurement (pp. 6-9).

Braš Roth, M., Gregurović, M., Markočić Dekanić, A. & Markuš, M. (2008) PISA 2006: Prirodoslovne kompetencije za život [PISA 2006: Science competencies for tomorrow’s life]. Zagreb: NCVVO:

Bridgeman, B. (2008) Experiences from large scale computer-based testing in the USA. Paper presented at International workshop The transition to computer-based assessment, Reykjavik, 29-30 September.

Kim, D.-H. & Huynh, H. (2008) Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course English test, Educational and Psychological Measurement, 68(4) 554-570.

Martin, R. (2008) New possibilities and challenges for assessment through the use of technology, in F. Scheuermann & A. Guimaraes Pereira (Eds) Towards a research agenda in computer-based assessment: Challenges and needs for European Educational Measurement (pp. 6-9).

Mead, A. D. & Drasgow, F. (1993) Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis, Psychological Bulletin, 113(3) 449-458.

Pitcher, N., Goldfinch, J. & Beevers, C. (2002). Aspects of computer-based assessment in mathematics, Active Learning in Higher Education, 3(2) 159-176.

Ripley, M. (2008) Technology in the service of 21st century learning and assessment, in F. Scheuermann & A. Guimaraes Pereira (Eds) Towards a research agenda in computer-based assessment: Challenges and needs for European Educational Measurement (pp. 22-29).

Wang, S., Jiao, H., Young, M. J., Brooks, T. & Olson, J. (2008) Comparability of computer-based and paper-and-pencil testing in K-12 reading assessment: A meta-analysis of testing mode effects, Educational and Psychological Measurement, 68(1) 5-24.

Wilhelm, O. & Schroeders, U. (2008) Computerized ability measurement: Some substantive Dos and Don’ts, in F. Scheuermann & A. Guimaraes Pereira (Eds) Towards a research agenda in computer-based assessment: Challenges and needs for European Educational Measurement (pp. 76-84).

Yeh, S. S. (2006) Reforming federal testing policy to support teaching and learning, Educational Policy, 20(3) 495-524.

The author:

Vesna Buško; University of Zagreb
Dep.of Psychology, Faculty of Humanities and Social Sciences
Ivana Lučića 3, 10000 Zagreb, Croatia
vbusko@ffzg.hr, vbusko@inet.hr

Vesna Buško is associate professor of psychology at the Faculty of Humanities and Social Sciences, University of Zagreb, where she chairs the Unit of Psychometrics. She teaches several subjects within the area of quantitative and psychometric methodology for undergraduate, graduate, and doctoral students at the Department of Psychology. Her main research interests include the assessment of cognitive abilities, study of emotional intelligence, and applications of structural equation modelling methodology. She is currently editor-in-chief of the Review of Psychology, international journal of the Croatian Psychological Association.

Comparing Paper-and-Pencil and Online Assessment of Reasoning Skills
A Pilot Study for Introducing Electronic Testing in Large-scale Assessment in Hungary

Benő Csapó, Gyöngyvér Molnár and Krisztina R. Tóth
Centre for Research on Learning and Instruction, University of Szeged
Research Group on the Development of Competencies, Hungarian Academy of Sciences

Abstract:
Computer-based assessment offers so many advantages that sooner or later it probably will replace paper-based testing in a number of areas. Primary and secondary schools are the settings where frequent and reliable feedback is most needed; therefore, recent work has focused on this area. Major international organizations and institutions (e.g., OECD PISA, ETS, NCES and CITO) are piloting the possibilities of transferring existing testing systems to the new medium and exploring new territories offered by recent technologies (see: Intel, Microsoft, and Cisco Education Taskforce 2008). However, the transition from paper-and-pencil (PP) to computer-based (CB) assessment raises several questions. Some of these are related to the availability of the necessary technological conditions at schools. Due to the rapid technological progress, production of energy and cost-efficient computers and the proliferation of new technologies in the schools of developed countries, these problems may be considered solved in a few years. The two test delivery media may affect different groups of participants in different ways and this concerns equity issues. Test administration mode affects participants' answering strategies as well (Johnson and Green, 2006). Furthermore, the perennial question of validity persists: what do computerized tests really measure? These questions are interrelated, and in order to make the transition process smother, they require careful analyses. This paper focuses on the test mode effect of assessment by using identical PP and CB tests and presents some early results of the first major pilot study of online testing carried out in public education in Hungary.

Background and context of the study

In general, there are two major groups of arguments for shifting the assessment from PP to CB: (1) existing tests can be administered by the means of technology more efficiently; therefore, well established assessment programs should also be transferred to the new medium, and (2) by means of technology (especially by exploiting the possibilities of multimedia) types of knowledge and skills can be measured that are not measurable (or cannot be measured so well) by means of PP tests. In the first approach, the comparability off PP and CB testing is crucial, whereas in the latter case there is no basis for comparison. Validity issues are equally important in both cases.

In our long-term project, we are going to apply a step-by-step approach, introducing new features offered by technology gradually, and controlling the effects caused by these new features. Therefore, first we transfer our existing PP tests into computer format, study how they work, treat the emerging problems and replace PP with CB assessment where possible.

In the first phase, we also try to identify areas of education, where computerized tests are most needed, and where the possible side effects may be minimized. Therefore, we intend to begin large-scale implementation of online assessment programs for formative and diagnostic assessment. Formative assessment can fulfil its promises only if it is frequent and feedback is immediate. In both respects, CB testing is more appropriate than PP. Formative assessment is even more efficient in an individualized educational context, where students follow their own developmental path. In such cases, students' results should be connected longitudinally, and it is also much easier if data are collected via computers. Furthermore, in the case of younger students, the "digital gap" is not too wide yet, and frequent computer usage may equalize their computer familiarity. A low-stake testing context and a basically helpful approach is expected to lower test anxiety and builds positive attitudes towards CB testing both in students and in their teachers. In the context of formative and diagnostic testing, security issues (e.g. preventing cheating) are less crucial; therefore, the implementation faces fewer constraints. Rare summative high-stake tests do not have all

these positive features; therefore, beginning the implementation of nationwide CB assessment with formative testing in the younger cohorts seems to be a logical decision (Csapó, 2008).
The study we report here is the first step along this road. The test we chose for the experiment has already been used in previous studies. It is not a formative one, but the data collection is part of a longitudinal project and the participants are relatively young.

Examining the differences between paper-and-pencil and computer-based testing

CB tests are often introduced without any piloting. This practice is based on the hidden assumption that PP and CB components should produce equivalent results if the content and cognitive activities of the two are identical (Clark, 1994). However, in most test mode effect studies significant differences have been found depending on the measured area (Clariana and Wallance, 2002).

In the USA, the possibilities of *Technology-Based Assessment* (TBA) were explored in the framework of a project carried out by the National Center for Education Statistics (NCES). The main aim of the study was to prepare to shift NAEP from paper bases to TBA. The project examined four main questions, (1) measurement, (2) equity, (3) efficiency, and (4) operational issues. The project covered three domains: Mathematics Online, Writing Online (results of these two domains were published in one volume, see Sandene, Horkay, Bennett, Allen, Braswell, Kaplan, and Oranje, 2005), and Problem Solving in Technology-Rich Environments (Bennett, Persky, Weiss and Jenkins, 2007). Testing Mathematics and Writing focused on the possibilities of delivering former PP tests online, while Problem Solving explored new test formats. Results suggested that, on average, CB testing worked well; however, media effect slightly depended on item format, and achievements on participants' computer familiarity.

In Europe, some countries have already implemented certain forms of computerized tests and several other national institution have been working on the introduction of large-scale computerized testing (see Scheuermann and Guimarães Pereira, 2008).

As for large international organizations, OECD is advancing an agenda of promoting the application of information-communication technologies (ICT) in education, including the application of technology in its flagship project, PISA. In the framework of ICT feasibility study information was gathered about a number of key developments regarding different ICT skills and their usability and delivery issues from technical and psychometric perspective (Lennon, Kirsch, Von Davier, Wagner, and Yamamoto, 2003). The first CB test administration within PISA took place in 2006 in the framework of Computer-Based Assessment of Science (CBAS). The results show the effect of the test delivery media; namely, the transition from PP to CB testing could lead to psychometric differences. Furthermore, the results confirm the importance of analyzing the validity issues by shifting from PP to CB testing. An important indicator of the difficulties is that out of the 57 participating countries, only three took part in CBAS. In PISA 2009, Electronic Reading Assessment (ERA) will represent computerized testing. More than twenty countries participated in the field trial of the instruments, but probably only 18 of them will do the main study. For the PISA 2012, Problem Solving is proposed as an international option, and it is planned to be assessed by the means of computers.

Depending on the usage of new features offered by technology, paper-and-pencil and computer-based testing may lead to different results, even if the same construct is to be measured. For this reason it is crucial to examine achievement differences at test, subtest, item, and subsample level across delivery media to identify items, item types, and subgroups of the sample behaving differently in PP and in CB mode and to confirm the key factors that relate to the test mode effect.

Objectives of the project and the pilot study

The purpose of the pilot study this paper reports on are: (1) to devise methods for analyzing differences between PPT and CBT; (2) to study the influence of the medium of assessment in a curriculum-independent competency field; (3) to investigate test/subtest/item level differences between PP and CB tests by focusing on validity issues; and, (4) to find an association between achievement differences and background variables. The main, long-term aim of the whole project is the preparation of a system-wide online measurement in Hungary.

Methods

Participants

Participants were fifth grader (11-years-old) primary school students drawn from a larger representative sample participating in a longitudinal project. The original longitudinal sample was composed in 2003. Over 5,000 students were chosen as a representative sample of the population entering schools. School classes were the unit of sampling; 206 classes out of 106 schools comprised the sample. A comprehensive school readiness test was administered to all students at the beginning of the project, and later several tests at the end of each school year. The main focus of the assessment has been mathematics and reading comprehension.

Because of several reasons (failing, changing schools, reorganization of schools etc.), the size of the sample decreased. At the end of the fifth year, there were 218 classes and 4,044 students in the longitudinal sample. Whole school classes were selected for the online assessment as well. Altogether, 68 classes from 34 schools and 843 students participated in the present study. 53% of the sampled students were boys.

Representativeness was not aimed for when composing the present sub-sample, but the deviation from a representative composition was controlled by the distribution of mothers' educational level. Table 1 compares the distribution of mothers' educational level of the original longitudinal sample and the sample of online assessment. According to the Chi-squared test results, the two distributions are the same, as they do not differ significantly. It means that the sample used in this study may be considered as representative for the entire fifth-grade school population of Hungary regarding mother's education, which is one of the most decisive background variables of students' developmental level.

	Representative sample	Pilot sample	$\chi2$	p
Below elementary school	2.8	1.4	7.13	.211
Elementary school	17.6	11.9		
Vocational school	28.2	30.7		
Matura examination	32.7	29.6		
BA degree	13.2	20.5		
MA degree	5.5	5.9		

Table 1: The distribution of mothers' educational level of the original longitudinal sample and the sample of online assessment

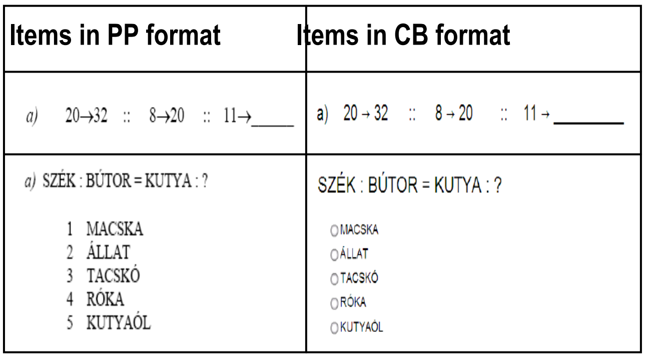
Instruments

In this study, students' inductive reasoning skills were assessed by the PP and CB version of a 58-item test in June 2008. Inductive reasoning was chosen, because it is considered a basic component of thinking, and is one of the most broadly studied construct of cognition (Csapó, 1997). The choice of inductive reasoning for this study was also supported by the assumption that in the case of such a general cognitive skill, no significant learning takes place between the two (PP and CB) testing sessions. (Results supported this assumption: no systematic improvement was assessed on the second test.)

The inductive reasoning test is comprised of three subtests: number analogies, number series and verbal analogies. The number analogy and number series subtests are composed of open-ended items, test takers are expected to answer them by giving (writing down in PP mode and typing via keyboard in CB mode) certain numbers. The verbal analogy items are multiple choice questions. The test is time-limited; participants have 35 minutes to complete the tests.

The whole inductive reasoning test was converted into a computerized version by preserving all features of the original one in order to make the two tests as similar as possible. However, the CB version of the test contained on the one hand a progress bar, which displayed where the students were in the test, and on the other hand, a back and next button for navigating in the test. In the multiple choice questions in PP format students had to circle the letter of the answer, in CB format they had to use radio button for giving their answer

(Figure 1). The computer-based version of the test has also a fixed-length form and it was delivered via the Internet.



Items in PP format	Items in CB format
a) 20→32 :: 8→20 :: 11→____	a) 20→32 :: 8→20 :: 11→______
a) SZÉK : BÚTOR = KUTYA : ? 1 MACSKA 2 ÁLLAT 3 TACSKÓ 4 RÓKA 5 KUTYAÓL	SZÉK : BÚTOR = KUTYA : ? ○MACSKA ○ÁLLAT ○TACSKÓ ○RÓKA ○KUTYAÓL

Figure 1: Layout of PP and CB items of number analogies and verbal analogies subtests

Procedures

First, all students took the inductive reasoning test in PP format. Then, a few weeks later the online version of the test was also administered to the same population. The PP test was taken in regular classrooms and the CBT version was taken in specially equipped computer rooms. The online data collection was carried out with the TAO platform.

To depict relationship between background variables and students' achievement, further information were collected from students as well as their teachers by means of questionnaires. The student questionnaires contained questions regarding students' computer familiarity, ICT-related attitudes and social background (gender, parents' education, school grades, subject attitudes, future plans). Teachers completed a follow-up questionnaire in an e-mail about the ICT equipment of schools, students' previous ICT experiences, and teachers' observations regarding the testing process to have feedback about the experience of using the online system.

For the delivery of the online tests and questionnaires, TAO (Testing Assisté par Ordinateur – Computer-Based Testing) was used (Plichart, Jadoul, Vandenabeele and Latour, 2004; Farcot, and Latour, 2008; Martin, 2008). TAO is an open source software developed by the Centre de Recherche Public Henri Tudor and the EMACS research unit of the University of Luxembourg.

Detailed item analyses were performed by using several means of classical test theory and IRT.

Results and discussion

The reliability index of the PP inductive reasoning test did not differ for the main longitudinal sample and for the sub-sample that took part in the online testing as well (Cronbach-α=.91). There was no significant difference between the reliability index of the PP test and CB test (Cronbach-α=.90 in the CB version) either.

Results of the present PP assessment of inductive reasoning for the entire fifth-grader longitudinal sample and the sub-sample participating in the online testing are compared in Table 2. Data show that the mean of the pilot study sub-sample was larger but the difference was not significant. The only significant achievement differences in PP mode was found in the verbal analogy subtest, where students in the pilot study got higher (p<.05) scores than students from the entire representative sample.

	Longitudinal sample		Pilot study (PP)		t	p
	Mean (%)	SD (%)	Mean (%)	SD (%)		
Inductive reasoning	26.8	14.8	27.2	14.9	.68	.49
Number series	14.3	11.1	14.3	11.5	-.06	.95
Verb analogy	38.5	21.1	40.3	21.5	2.15	.03
Number analogy	27.5	22.3	27.0	21.6	-.66	.51

Table 2. Test and subtest-level results for the entire fifth-grader longitudinal sample and the pilot study

The average scores on PP test and the online test differ significantly; however, there is only a minor difference (see Table 3). Students' achievement was higher in PP mode than in CB mode with one exception, in the field of verbal analogy where the subtest contained multiple-choice items, students achieved higher scores in CB than in PP mode. The highest media effect was noticeable on open ended items requiring calculations (number series items).

			Mean	SD	T	p
CBT total	-	PPT total	-1.17	9.48	3.57	0.000
CBT Number analogy	-	PPT Number analogy	-1.62	17.53	2.68	0.007
CBT Verbal analogy	-	PPT Verbal analogy	2.50	13.78	-5.27	0.000
CBT Number eries	-	PPT Number series	-4.38	11.88	10.71	0.000

Table 3: Comparing test and subtest level achievements in PP and CB mode

A strong correlation (r=.79) is found between the total scores of the two versions of the test. As for the subtests, there are differences between the strengths of correlations: the weakest relationship is between the different versions of the number series tests (r=.62 and .42, respectively), whereas the strongest one characterizes the test of verbal analogies (r=.80).

Regarding gender analyses, there were no achievement differences between the achievement of boys and girls in PP or in CB test results (Table 4). On subtest level, when an analysis of the results in the two media took place separately, gender differences were found only in PP testing on the verbal analogy subtest. Comparing the PP and CB results by gender, several differences are noticeable across delivery media. Girls achieved significantly better on the PP test than on the computerized version (m_{PP}=27.66 and m_{CB}= 26.04); however, the delivery media had no significant impact on boys' achievement at test level (m_{PP}=26.73 and m_{CB}=25.99). At subtest level, girls performed on significantly different levels in every subtest regarding delivery media, whereas boys had significant differences in mean scores on two subtests (number series and verbal analogy).

	PP	CB	Girls	Boys
	Between gender		Within gender	
Inductive reasoning	n.s.	n.s.	PP>CB; p<.05	n.s.
Number analogy	n.s.	n.s.	PP>CB; p<.05	n.s.
Verbal analogy	girls>boys; p<.05	n.s.	CB>PP; p<.05	CB>PP; p<.05
Number series	n.s.	n.s.	PP>CB; p<.05	PP>CB; p<.05

Table 4: Comparing gender differences between and within gender according PP and CB results

To confirm some key factors relating to the test mode effect, ICT expertise and ICT familiarity also differ between genders. Boys have more expertise in computer usage than girls, but there are several side effects that require further analyses.

In sum, this study revealed that the basic conditions for online testing are available in average Hungarian schools. Tests may be delivered via the Internet without major obstacles. Current technical conditions experienced in the schools taking part in the study seem to be sufficient for low-stake (formative, diagnostic) testing. In order to ensure the conditions for high-stake testing, students should be more equally exposed to computer experiences. If the equivalence of PP and CB testing is a requirement, test items should be carefully analyzed. For high-stake testing, further technological development is necessary to ensure standardized testing condition.

This work confirmed that probably there will be a shorter or longer period when paper-based and computer-based testing co-exist. It turned out that there are visible expectations to relate this new type of assessment to the existing ones. Both students and teachers are interested to know the differences. Decision makers also would like to see how this new instruments may fit the purposes of monitoring changes in education generated by reform attempts. All these expectations require further analysis of the media effects in a framework that goes beyond the pure technicalities.

The study has also revealed that large-scale CBT may be completed in Hungarian schools, in a country where several programmes helped schools to get equipped with basic ICT facilities, but no specific attention was paid to ensuring the conditions necessary for online assessments. On the one hand, this piloting work allows the establishment of standards concerning the minimum equipment in schools required to participate in online assessments and provides decision makers with a further frame of reference when designing the completion of equipping schools with ICT facilities. On the other hand, experimental work can be further progressed; it is not hindered by the current technological constrains. The accumulated experiences at the area of online testing may have a positive effect on the technological improvement as well.

Several issues have also been raised during the piloting work that is worth further analysis and imply developmental work beyond technical issues. Implementing a successful assessment system requires that students, teachers, parents and stakeholders in general, accept the results produced by the measuring instruments. In general, they have to trust in the system. Computerized assessment adds further unfamiliar elements to the already complex assessment processes. Designing a feedback system with rich explanations, familiarizing students with the system, training teachers, informing stakeholders should also be kept in the horizon of the developmental programs. These observations indicate the complexity of social aspects of introducing innovative assessment technologies. Taking into account the divergence between the rapid technological development and the time-frame that is necessary for their educational implementation, experimenting with the most advanced technology based assessment should also be launched as early as possible, even if their large-scale, system level application cannot be expected in the near future.

References

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. Research and Development Series (NCES 2007–466). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Clariana, R. & Wallance, P. (2002). Paper-based versus computer-based assessment: key factors associated with test mode effect. British Journal of Educational Technology, 33 (5) 593-602.

Clark, R. E. (1994). Media will never influence learning. Educational Technology Research and Development, 42(2) 21-29.

Csapó, B. (1997): The Development of Inductive Reasoning: Cross-sectional Assessments in an Educational Context. International Journal of Behavioral Development, 20(4), 609–626.

Csapó, B. (2008). Integrating recent developments in educational evaluation: formative, longitudinal and online assessments. Keynote Lecture. The European Conference on Educational Research. Gothenburg, Sweden. 8-9 September 2008.

Farcot, M. & Latour, T. (2008). An open source and large-scale computer-based assessment platform : A real winner. In F. Scheuermann, & A. Guimarães Pereira (Eds.), Towards a research agenda on computer-based assessment: Challenges and needs for European Educational Measurement (pp. 64-67). Ispra: European Commission Joint Research Centre.

Intel, Microsoft, and Cisco Education Taskforce (2008). Transforming education assessment: Teaching and testing the skills needed in the 21st century. A call to action. Unpublished manuscript.

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). 1-34.

Lennon, M., Kirsch, I., Von Davier, M., Wagner, M., & Yamamoto, K. (2003). Feasibility study for the PISA ICT literacy assessment. Princeton, NJ: Educational Testing Service.

Martin, R. (2008). New possibilities and challenges for assessment through the use of technology. In F. Scheuermann, & A. Guimarães Pereira (Eds.), Towards a research agenda on computer-based assessment: Challenges and needs for European Educational Measurement (pp. 6-9). Ispra: European Commission Joint Research Centre.

Plichart P., Jadoul R., Vandenabeele L., Latour Th. (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, Luxembourg.

Scheuermann, F. & Guimarães Pereira, A. (2008). (Eds.) Towards a research agenda on computer-based assessment: Challenges and needs for European Educational Measurement. Ispra: European Commission Joint Research Centre.

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project, Research and Development Series (NCES 2005–457). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

The authors:

Benő Csapó, Gyöngyvér Molnár, Krisztina R. Tóth
Institute of Education
University of Szeged
Petőfi sgt. 30-34.
H-6722 Szeged, Hungary

E-mail: csapo@edpsy.u-szeged.hu
gymolnar@edpsy.u-szeged.hu
tothkr@inf.u-szeged.hu

Benő Csapó is a Professor of Education at the University of Szeged and the head of Research Group on the Development of Competencies, Hungarian Academy of Sciences. His research interests include cognitive development, organization of knowledge, educational evaluation, longitudinal studies in education, and assessment methods.

Gyöngyvér Molnár is an Associate Professor of Education at the University of Szeged. Her main research areas are complex problem solving, application of information-communication technologies in education, educational measurement, especially issues of Item Response Theory.

Krisztina R. Tóth is a Ph.D. student at the Graduate School of Educational Sciences, University of Szeged. She studies the implementation of computerized assessment.

..................................IV. Methodologies of Computer-based Testing

Computerized and Adaptive Testing in Educational Assessment

Nathan A. Thompson & David J. Weiss
Assessment Systems Corporation & University of Minnesota, USA

Summary
A keynote-style overview of the issues involved in computerized delivery of educational assessments is provided. Advantages and disadvantages of different approaches are reviewed, and the technical advantages of the sophisticated technology of algorithmic testing approaches, including computerized adaptive testing and computerized classification testing are explored further. Problems with the use of the Internet to deliver these types of test are briefly addressed.

The application of personal computers has introduced many advantages to the large-scale assessment of students. However, not all computerized tests are created equal. The utilization of computers for student assessment can vary substantially, and likewise the issues and advantages that accompany each form of utilization can differ substantially. This paper outlines the methods of delivering computerized tests, discusses the advantages and disadvantages, and provides an introduction to the algorithms that are applied.

At the most basic level, computerized assessments can be categorized into those that are locally controlled and those that are remotely controlled. Locally controlled assessments are those where the test delivery engine is local with respect to the student, either directly on the student's computer or on a local area network (LAN) with minimal time lag. Remotely controlled assessments are those where the testing engine resides on a server that can be hundreds or even thousands of kilometres away. The fundamental psychometric difference between these approaches is the fact that locally controlled tests are much more likely to deliver standard assessments, partly because of time lag issue. Because the goal of assessment is to reduce the amount of construct-irrelevant variance as much as possible, the ability to standardize the delivery of a test is of utmost importance. Consequently, most of this paper will focus on locally-delivered assessments.

A more common way to categorize assessments is by the algorithms that underlie the delivery engine. Computerized fixed-form tests (CFT) — also referred to as conventional, linear or traditional tests — generally deliver a predetermined set of items to the student. This is equivalent to paper-and-pencil testing (PPT), but with the items delivered to the student on a computer screen rather than in a paper booklet. Variable-form testing approaches utilize the computing power and interactive ability of a computer to administer a set of items that is determined at examination time, rather than a predetermined set of items. Two widespread variable-form approaches are computerized adaptive testing (CAT; Weiss & Kingsbury, 1984) and linear-on-the-fly testing (LOFT; Luecht, 2005), also referred to as automated test assembly (ATA; Lin, 2008).

Advantages of Computerized Testing

CFT, which is the least sophisticated of computerized assessments, still provides advantages over PPT. Obviously, the need for printing, storage, and distribution of booklets, as well as the collection and scanning of answer sheets, is no longer applicable. This in turn leads to the advantage that CFT is able to make use of item formats not available with PPT, such as multimedia stimuli. Similarly, the computer is able to record certain information not available in PPT, such as item response times. Further, tests can be continuously available rather than administered in time windows constrained by the logistical issues typically involved with printed forms. An additional advantage is that item sequences can be randomly scrambled for each student, increasing security. However, one of the most recognized advantages is that results are immediately available, either for the student or the teacher.

There are, of course, some disadvantages to CFT as compared to PPT. The most practical of these is the fact that PPT might simply be more economical for smaller testing programs. A further concern is that some testing populations might be uncomfortable with computers. However, both of these disadvantages do not reflect a problem with the test itself, but rather aspects of the testing program, so little can be done by psychometricians or test developers to address these concerns.

CFT also has distinct disadvantages as compared to variable-form testing. The use of fixed forms greatly limits the number of item sets that are possible for each student. This can lead to item exposure problems and consequent potential item sharing between students who have taken a test and those yet to take it. Additionally, time and items are wasted by the administration of items to a given student that are too difficult or too easy to be of psychometric value. This also contributes to item exposure issues.

The important facet of these issues, however, is that they are issues that are endemic to the test itself, and the application of computers presents the opportunity to address the issues psychometrically. This is one reason for the application of variable-form testing approaches: the test can be designed to specifically construct a test for each student to control for these possible problems.

There are three primary variable-form approaches: LOFT, CAT, and multistage testing. LOFT has more in common with CFT than the other two, as it uses a fixed-length set of items administered to a student in a predetermined sequence. What sets it apart is that a new form is constructed at examination time for each student; each form is designed to equivalent to the extent that the test sponsor desires. The primary advantage of this approach is the possibility of a very large number of forms, which obviously increases the security of the test.

Multistage tests are tests that are administered in mini-forms or “testlets,” which are routed to make more efficient use of the items. For example, rather than administer a total form of 50 items, a set of 20 items could be administered first, and then a set of 30 items administered. The difficulty of the second set depends on the student’s performance on the first set; students who perform well will receive more difficult items. This approach has far fewer possible forms than LOFT, therefore possibly creating fewer security issues, but makes an effort to tailor the test so that students do not receive as many items of inappropriate difficulty.

The most sophisticated type of variable-form testing is an algorithmic approach, where the test is designed to be administered with a dynamic, interactive algorithm. This is in contrast to multistage testing, where there are fixed routes between the testlets. Instead, the algorithm will adapt the test to each student, not just with respect to which items are selected but also with respect to how many items are selected. If the goals of the test are satisfied after only 10 items, then the student’s exam can be concluded. This is important because not only does the test make better use of items, it does not administer any more items than necessary, which can greatly reduce item exposure.

The most well-known type of algorithmic testing is CAT, which is a test where the algorithm is designed to provide an accurate point estimation of individual ability or achievement. A similar, but lesser-known, approach is computerized classification testing (CCT), also known as sequential testing, where the algorithm is designed to classify students. For example, students can be classified as pass/fail or into educational proficiency levels such as basic/proficient/advanced.

The algorithmic approach realizes several new advantages in addition to those previously discussed. The most important advantage is due to the variable-length aspect of CAT/CCT; typically only about half as many items are needed to provide precision equivalent to conventional tests — in some cases considerably fewer. Over large numbers of students, this can amount to substantial savings in test seat time, while at an individual level it allows for more time to be spent giving feedback and instruction to students.

Additionally, a properly designed and implemented CAT can affect the motivation of students. Because lower-ability students will receive easier items, they will become less discouraged and stressed. Conversely, high-ability students will not be wasting time on items that are far too easy; they will receive items that appropriately challenge their high ability.

A further advantage of CAT or CCT is that the degree of score precision or classification accuracy can be specifically controlled in the aggregate. For example, a CCT can be designed to produce fewer than 1% errors of misclassification, assuming that the item bank is of high enough quality. Similarly, a CAT can be designed to provide an equivalent level of psychometric precision for each student, something that is extremely difficult to do with fixed tests forms. This results in a new conceptualization of “test fairness.”

CAT can also be modified to efficiently and effectively measure individual change or growth to monitor student progress (Kim-Kang & Weiss, 2008). Although the vast majority of testing applications only consider one point in time, educational assessments are often concerned with student progress, making CATs designed specifically to measure individual change very appropriate to that application.

To further explore the advantages offered by CAT and CCT, the technical aspects of each are defined. However, because the technical aspects are usually based in item response theory (IRT), a brief introduction to IRT is provided first.

IRT

IRT is a psychometric theory that is based on the premise that the probability of a correct response to an item is a function of a student's ability (denoted by θ), and this item response function can be approximated by a cumulative normal curve or a logistic approximation to it. IRT enables advanced methods of CAT by placing students and items on the same scale.

The item response function (IRF) is the backbone of IRT. For educational assessment where items can differ in discrimination strength and guessing often plays a role, an appropriate form of this function is the three-parameter logistic model (3PL; Hambleton & Swaminathan, 1985, Eq. 3.3):

$$P_i(X_i = 1 | \theta_j) = c_i + (1 - c_i)\frac{\exp[Da_i(\theta_j - b_i)]}{1 + \exp[Da_i(\theta_j - b_i)]} \qquad (1)$$

where

a_i is the item discrimination parameter,

b_i is the item difficulty or location parameter,

c_i is the lower asymptote, or pseudoguessing parameter,

D is a scaling constant equal to 1.702 or 1.0.

A depiction of a 3PL IRF is shown in Figure 1. The x-axis represents the scale that references both student ability (θ) and item difficulty, while the y-axis is the probability of a correct response, from 0.0 to 1.0. The item discrimination parameter a_i determines the slope of the curve at its steepest point, where a steeper curve indicates a better-discriminating item. The difficulty of the item is represented by the point on the θ scale where the slope is the steepest; an item located farther to the right is more difficult. Lastly, the c_i parameter is the lower asymptote; for example, with a multiple-choice item of 4 options, even the lowest ability student has a 25% chance of guessing the correct answer.

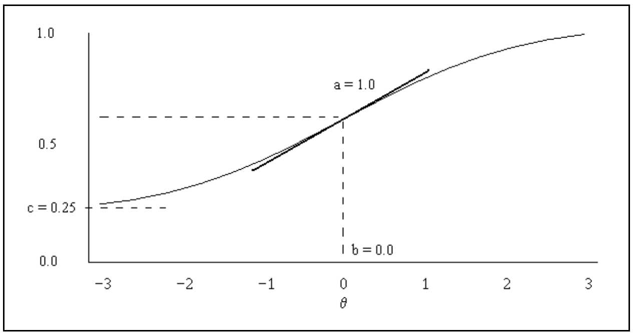


Figure 1: Depiction of a 3PL IRF

If a student has answered items, and the items have been scored as correct or incorrect, the IRF for each correct item and (1–IRF) for each incorrect item are multiplied. This produces a curve that is called the *likelihood function*. The highest point of this function is the *maximum likelihood ability estimate* (MLE), as it represents the point on θ that is the most likely to be the student's ability given the pattern of item responses. The precision of this estimate is termed the *standard error of measurement (*SEM); a higher value indicates less precision. The SEMs for a student can also be obtained from the likelihood function based on how "peaked" the likelihood function is.

Figure 2 presents two example likelihood functions. The curve on the left has a maximum near ˜ 0.6, while the curve on the right has a maximum of 2.2. These points on the x-axis then represent the MLEs for the two students.

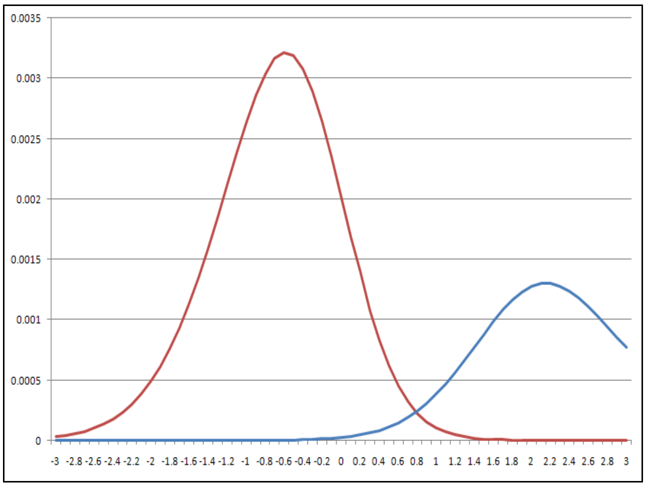


Figure 2: Two example likelihood functions

An Introduction to CAT

A CAT consists of five technical components:

1. An item bank calibrated with a psychometric model (e.g., the 3PL).
2. A starting point on the θ scale for a student.
3. An item selection algorithm.
4. A scoring procedure.
5. A termination criterion.

The basic CAT algorithm works by first specifying components 1 and 2 for a given student, then cycling through components 3, 4, and 5 until the termination criterion is satisfied, at which point the test is terminated.

While it is possible to design a CAT based on classical test theory, most operational CATs make use of IRT as the psychometric model. The remaining components then have IRT-based definitions. The starting point can be randomly selected within a small interval of θ, or fixed to the mean of the population (0.0 if the scale is determined by the population). Items are selected by maximizing their Fisher information (Weiss & Kingsbury, 1984), which is a function of the item parameters. The scoring procedure refers to the MLE or a related method. Finally, the termination criterion is often determined as a minimum value of SEM.

Figure 3 presents a flowchart of the CAT process. An item bank is constructed for the CAT, and each student starts their test at a certain value of θ. An item is selected from the bank based on that θ value, delivered to the student, and then scored. The student's θ is re-estimated based on that new piece of information, and if the termination criterion is satisfied, the test is concluded with that estimate of θ. If it is not, the process cycles back to the selection of another item.

This flowchart also presents some of the practical constraints involved in CAT. For example, selection of the next item is often subject to content domain targets, item exposure levels, randomization schemes, and, of course, whether the item has already been administered to the student.

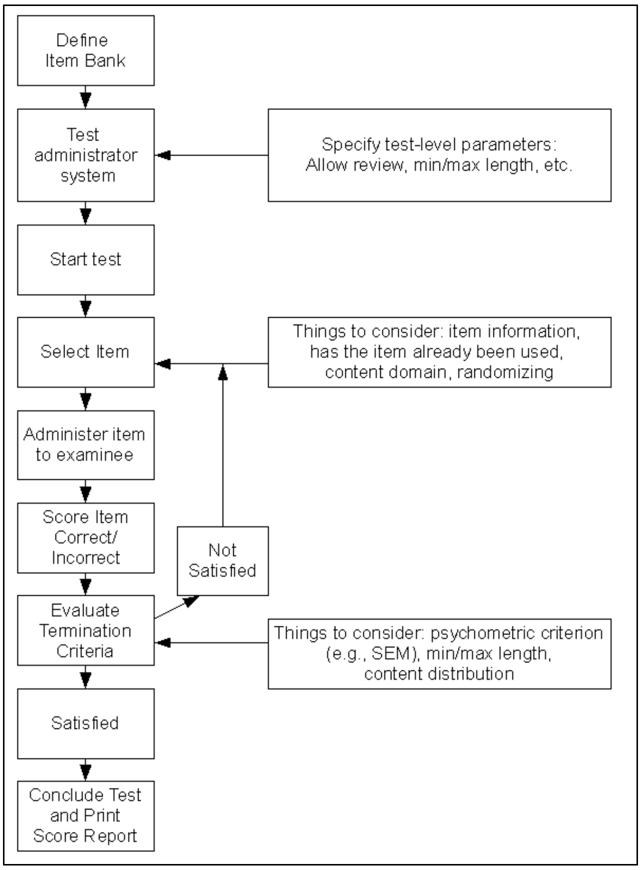

Define Item Bank
Test administrator system
Specify test-level parameters: Allow review, min/max length, etc.
Start test
Select Item
Things to consider: item information, has the item already been used, content domain, randomizing
Administer item to examinee
Score Item Correct/ Incorrect
Not Satisfied
Evaluate Termination Criteria
Things to consider: psychometric criterion (e.g., SEM), min/max length, content distribution
Satisfied
Conclude Test and Print Score Report

Figure 3: Flowchart of CAT Algorithm

Figure 4 presents an item-by-item depiction of the θestimation process in CAT, with the x-axis representing the point in the test. The box in the center of each band is the MLE, while the band is the SEM added and subtracted from the MLE. In this example, there is no score at item 0, as it is still θ= 0.0. After the first item (which was answered correctly), the MLE is updated to θ= 0.5 (based on a correct answer to item 2), and after that to θ = 1.0 (again based on a correct answer). Eventually, the iterative process "zeroes in" at an estimate of approximately θ = 0.80. Note that the error bands generally decrease in width as each item is answered and scored.

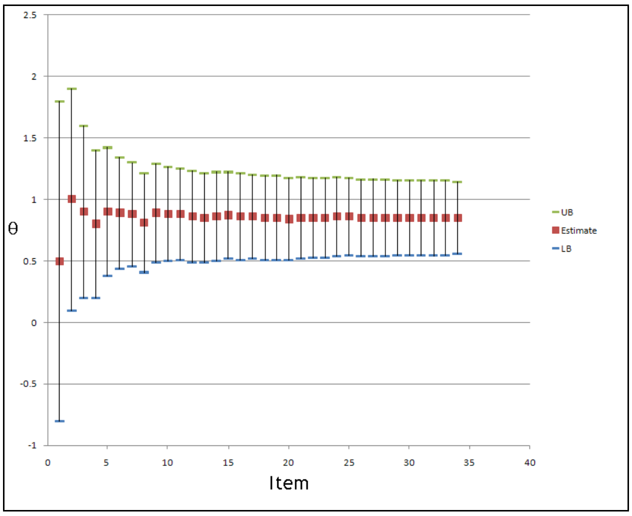


Figure 4: Item-by-item view of CAT

Implementing CAT

There are three stages to an implementation of a CAT approach to a testing program that has an established fixed-form test. The first step is to evaluate the CAT approach by administering the fixed-form in an adaptive manner; each student still receives the same set of available items, but each item is dynamically selected with the algorithm. All items on the form are still administered after the termination criterion is satisfied so that full-form θ estimates can still be determined. The reduction in test length can then be evaluated by determining a shorter ACT that correlates highly with the full test. The test length reduction can be from 50% upward to 95%, depending on the item bank quality and the length of the fixed form, and still produce θ estimates that correlate above r = 0.90 with the full-form estimates.

Fortunately, it is often not necessary to complete this stage by actually putting such a test out in the field to see if the reduction in test length is as large as desired. A research approach termed *post-hoc simulation* (Weiss, 2005) is designed to mimic this kind of study by simply taking past results from the fixed-form test and simulating how CATs would function for each of the students. The resulting CAT estimate is then easily compared to the full-form estimate actually obtained by each student.

The second stage is the use of partially adaptive tests, such as modified multistage tests. Specifically structured banks and sub-banks can be used to administer variable-length multistage tests to students. This provides an increase in efficiency by reducing test length. However, problems with this approach include inefficient use of item banks, misrouting, the fact that test length is still somewhat fixed, and lack of control over measurement precision. Often, this stage is optional, as the results of the research in the first stage can provide evidence that fully adaptive CATs will provide satisfactory results even with a given item bank.

The third stage is that of fully adaptive tests, where each individual item is selected dynamically, test length is completely variable, and measurement precision is firmly controlled. At this stage, the CAT algorithm as described previously is allowed to act as designed, within optional practical constraints such as item exposure and content distribution.

Testing for Classification

CAT can also be adapted for the classification testing. A confidence interval is obtained by multiplying the SEM by an appropriate normal deviate z (e.g., 1.96 for 95% accuracy), which is then added to and subtracted from the θ estimate. If this interval falls completely above a specified cutscore, the student is classified as above the cutscore, and vice-versa. If the interval contains the cutscore, another item is administered and the interval updated. Other than this modification to the termination criterion, the CAT algorithm remains the same.

This approach was initially suggested for CCT (Kingsbury & Weiss, 1979). However, further research determined that it was more efficient do specifically design the algorithms with respect to classification rather than adapting the point-estimation CAT approach (Reckase, 1983; Spray & Reckase, 1996). There are two primary differences in the CCT algorithm: the items are selected with respect to the cutscore and not the current θ estimate, and the scoring procedure and termination criterion are combined and utilize a different approach, the likelihood ratio. In fact, the estimation of an actual score on the θ scale is no longer necessary.

While the cutscore-based item selection concept is straightforward, the concept of a likelihood ratio is more complex. As the term would suggest, it is a ratio that utilizes the likelihood function. Two points are selected, and the values of each on the y-axis are compared in a ratio form. The test is terminated with the value of the ratio exceeds boundaries that are determined by user-specified nominal error rates, α and β (Wald, 1947):

Lower decision point = $B = \beta / (1 - \alpha)$ (2)

Upper decision point = $A = (1 - \beta)/\alpha$. (3)

For 95% nominal accuracy, which entails $\alpha = \beta = 0.025$, these result in the values 0.026 and 39.

There are two methods that have been used to select points on the likelihood function. Originally, it was suggested that they be fixed, where θ_1 was a point below the cutscore and θ_2 was a point above the cutscore (Reckase, 1983). However, it has been shown that it is more efficient to allow them to vary by selecting the highest points in predetermined regions of θ above and below the cutscore (Thompson, 2008).

CCT is easily extendable to multiple classifications, such as the educational proficiency levels basic/proficient/advanced. Unfortunately, tests with three or four cutscores are very difficult to do accurately. Additionally, this can affect the accuracy of summary percentages of students at each level (Betebenner et al., 2008).

Conclusions

Computerized test delivery offers significant advantages over PPT for most testing programs. Computerized tests can assess many things that paper-and-pencil tests are not able to assess. They can do it more quickly, more efficiently, and with lessened logistical burden. But most importantly, the interactive nature of the computer can be utilized in the delivery of the test to make better use of both item banks and student time.

An important issue in computerized test delivery is the administration of tests over the Internet. As previously mentioned, the most important aspect of test delivery is standardization to eliminate sources of construct-irrelevant variance. Unfortunately, current Internet speeds and traffic preclude the use of Internet-delivered CAT/CCT to deliver acceptably standardized tests. The reason for this is the need for calculation that must take place in the delivery engine. After a student answers an item, the response is sent back to the engine, compared to the item key to determine if it is correct, the MLE must be recalculated with the IRF of the most recent item, a new item is selected based on the new MLE and other constraints, and the new item is then sent back to the student's computer. If this process takes place on the student's computer or even in the next room, the time delay is negligible. However, if the process takes place hundreds or thousands of kilometers away, it is impossible to completely control the time delay. Some items might be presented quickly, while some might take 30 seconds.

There should be substantial concern for this issue, because while possible Internet speed increases with the widespread application of DSL, cable, T1 lines, and the like, the amount of traffic on the Internet is also increasing. For example, many more people read the newspaper online in 2008 than occurred even five years ago. It is this reason that it is impossible to control the effect of traffic on Internet speed and therefore test delivery.

Such effects might be acceptable in many testing situations with low stakes. A self-administered personality quiz, a customer satisfaction survey, or a summary test for an employee training project at a small company might have very few consequences if there is time lag between items. But educational and professional assessments, with stakes such as school graduation and professional licensure, face possible litigation if there are Internet delivery issues. Therefore, it is irresponsible from a psychometric and legal defensibility point of view to administer such tests via the Internet.

Yet computer-delivered tests remain the future of educational, psychological, and professional testing. The advantages that they offer are too important to ignore, especially from a business case perspective. In many applications, computerized tests can be delivered with a reduction in cost and an increase in precision and efficiency. CAT and CCT have the additional advantage of reducing individual student seat time, which allows for more time to be spent utilizing the feedback of the test to promote further instruction. As quality instruction is the goal of educational programs, the fact that CAT actively contributes to this goal gives it a central place in the future of educational assessment.

References

Betebenner, D. W., Shang, Y., Xiang, Y., Zhao, Y., & Yue, X. (2008). The impact of performance misclassification on the accuracy and precision of percent at performance level measures. Journal of Educational Measurement, 45(2), 119-137.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Norwell, MA: Kluwer Academic Publishers.

Kim-Kang, G., & Weiss, D.J. (2008). Adaptive measurement of individual change. Zeitschrift fur Psychologie / Journal of Psychology, 216, 49-58.

Kingsbury, G.G., & Weiss, D.J. (1979). An adaptive testing strategy for mastery decisions. (Research Report 79-5). Minneapolis, MN: University of Minnesota Psychometric Methods Program..

Lin, C-J. (2008) Comparisons between classical test theory and item response theory in automated assembly of parallel test forms. Journal of Technology, Learning, and Assessment, 6(8). Available online at http://escholarship.bc.edu/jtla/vol6/8/.

Luecht, R. M. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. Journal of Applied Testing Technology, 7(4) . http://www.testpublishers.org/Documents/JATT2005_rev_Criteria4CBT_RMLuecht_Apr2005.pdf.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), New horizons in testing: Latent trait theory and computerized adaptive testing (pp. 237-254). New York: Academic Press.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. Journal of Educational & Behavioral Statistics, 21, 405-414.

Thompson, N.A. (2008). Computerized classification testing with composite hypotheses. Under revision.

Weiss, D. J. (2005). Manual for POSTSIM: Post-hoc simulation of computerized adaptive testing. Version 2.0. St. Paul MN: Assessment Systems Corporation.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21, 361-375.

The authors:

Contact person:
Nathan A. Thompson, Ph.D., CCOA
Assessment Systems Corporation
2233 University Avenue, Suite 200
Saint Paul, MN 55114

E-Mail: nthompson@assess.com

Nathan A. Thompson, Ph.D., is the Vice President of Assessment Systems Corporation, where he oversees consulting and research work. His primary research interest is algorithmic testing approaches for computerized classification testing. Dr. Thompson received his Ph.D. from the University of Minnesota.

David J. Weiss, Ph.D., is a Professor at the University of Minnesota, where he established the Psychometric Methods program and served as its director for more than 30 years. He is co-founder and President of Assessment Systems Corporation. He also is the founding editor of the prestigious journal Applied Psychological Measurement and Editor for its first 25 years. Dr. Weiss is a pioneer researcher on computerized adaptive testing and the item response theory (IRT) that supports it. Dr. Weiss received his Ph.D. from the University of Minnesota in 1963.

Computerized Adaptive Testing of Arithmetic at the Entrance of Primary School Teacher Training College (WISCAT-pabo)

Theo J.H.M. Eggen & Gerard J.J.M. Straetmans
Cito, Netherlands & University of Twente, Netherlands

Abstract:
A few years ago the Dutch were shocked by the news that first-year students of primary school teacher training colleges were very poor in arithmetic. In response to this news "The Netherlands Association of Universities of Applied Sciences" (HBO-raad) ordered the implementation of a nation-wide, obligatory test for measuring prerequisite knowledge and skills in arithmetic. Unfortunately, a number of problems were involved that made the construction of this test particularly awkward. That is why it was decided to build a computerized-adaptive test package. Adaptive testing is a method for administering tests that combines computer technology with modern measurement theory to increase the efficiency of the testing process. Besides increased efficiency, CAT offers several other appealing advantages. In this article the concept of adaptive testing and the way it works are explained. In addition, information is presented about the test results and the experiences of the candidates during the first year of operation.

Some thirty years ago one of the authors of this article studied at a primary school teacher training college. He noted down the following recollection: "The training then was only three years of study instead of four, with scope for specialization in the final year. I chose arithmetic because I thought that it was a pretty important skill for a teacher to have. I remember how surprised I was that so few students made the same choice, all the more because the intervision meetings kept on showing that students were experiencing trouble teaching arithmetic because of their own lack of mathematical skills." In the decades that followed, the problems due to a lack of arithmetic skills were also noticed outside the training setting and the authorities kept imposing new measures to combat them. Since the beginning of 2006, we have known for a fact that those measures did not really help, because that is when research results were published showing that more than half of the first-year PABO students had insufficient arithmetic skills (Straetmans & Eggen, 2005).(*PABO = Pedagogische Academie Basisonderwijs = primary school teacher training college*) The national discussion that ensued, with often unsubtle newspaper headlines fanning the flames by suggesting that there were problems not only with the arithmetic skills of the PABO students but also with the Dutch education system as a whole, resulted in more decisiveness from the decision-makers. The *HBO-raad* (Netherlands Association of Universities of Applied Sciences) decided to develop an obligatory standardized arithmetic skills test that would be used as the basis for a binding recommendation on further study at the end of the first year of the course. The education minister ordered investigations into the causes of the lack of arithmetic skills among first-year PABO students so that a definitive solution could be obtained.

Then Cito developed a testing package that could be used to determine the arithmetic skills of first-year PABO students nationwide. This article begins by looking at the specific problems of making such measurements, followed by the principle of adaptive testing, the structure and algorithm for the testing package, the initial test results and some of the users' experiences.

Difficulties in measurement

Outsiders are often incredulous to hear that it takes professional testing experts months to develop what to them is just a simple little test for a particular school subject. People are generally insufficiently aware of the obstacles that can loom up on the way if you want to take crucially important decisions about people on the basis of a test result. However, anyone who reads the following statements by Suen (1990, pp. 5-8) will soon change their mind. "*The purpose of educational measurement is to describe people's characteristics as numerical scores. At first glance it is a deceptively simple task that anyone could do. However, when drawing conclusions about the candidates, it is too often assumed to be obvious that the scaling procedure (transforming the component responses into a score) is correct, that the observed score is a reliable reflection of the true score, and that the true score is in turn a truthful* *reflection of the quantity of the characteristic being measured."*

Suen's statements apply to any test that has to be used as the basis for serious decisions about candidates and it makes the construction of these types of tests a labour-intensive job. However, in the specific case of the PABO arithmetic test, a number of problems were involved that made constructing the test particularly awkward.

1. *Time of sitting.* In many examinations, all candidates are tested at the same time. The major benefit of this is that all the candidates take the same test items. However, no fixed dates have been defined for sitting the PABO arithmetic test. The training course, and indeed usually the individual arithmetic trainer, decides when the students are to be tested. This obviously creates a 'secrecy problem'. Where the consequences of the testing are major, candidates may also try unauthorized methods of obtaining a sufficient mark. Passing items on from one person to the next is one obvious and virtually uncontrollable response among candidates who are tested successively rather than simultaneously. This problem can only be tackled appropriately if there are enough variants of the test.
2. *Comparability of performance in the test.* The need for multiple versions of the test immediately introduces a new problem, namely that performance on the various test variants may not necessarily be comparable. A score of 8 correctly answered items out of 15 on test version A may indicate a different ability level than the same score on test version B, if the two tests differ in terms of difficulty. The level of difficulty of a test depends on the items of which it is composed. If numerous test versions have to be made, as is the case for the PABO arithmetic test, then more items are often required in these 'parallel tests' (as they are known) than can be constructed within a given timeframe and budget.
3. *Large variations in arithmetic skills.* The population of first-year PABO students is highly heterogeneous in terms of arithmetic skills, due to the differences in previous education. The majority of students have a HAVO diploma (*general secondary education*), but increasing numbers of students come with an MBO diploma (*vocational secondary education*). In addition, there are also people with a VWO diploma (*general secondary/pre-university education*). Studies have shown that there are substantial differences between the arithmetic skills of these groups (Straetmans & Eggen, 2005). This is awkward to deal with when constructing the tests. In order to make an accurate measurement, the level of difficulty of a test must be appropriate for the skills of the candidate to be assessed. However, if those skills vary widely, the test developer has no suitable target level point for determining the level of difficulty of the test (and therefore of the items to be administered in the test).

The solution: CAT

The solution for the above-mentioned problems was found in a special application of computerized testing that the literature generally refers to as CAT (computerized adaptive testing). Unlike other forms of computer-controlled testing, CAT is not merely about using a computer screen to take a (predefined) test and automatically processing the responses. Instead, it is (primarily) about the automated construction of a test from an item bank. There are basically three ways in which software can compose tests from an item bank. In the first method, the items are picked from the database using a random mechanism. Apart from setting the length of the test, you have no control at all over the test that the software produces. In the second method, the software attempts to stick within a test grid that prescribes how many items the test must include about each of the topics taught. The third method attempts to select items that will reduce the measurement error in the test score as far as possible. This is achieved by selecting items for which the level of difficulty is matched as well as possible to the skills of the person being tested. This third method is the approach adopted by adaptive testing.

The principle of adaptive testing is not new. It is also often used in oral examinations. If it appears that the item being asked was too difficult or too easy, a sensible examiner will then ask a simpler or more complex item as appropriate. The reason for this is that he will not learn much about the ability of the candidate if he just keeps on asking items that are too hard or too easy (Wainer, 2000). What *is* new is the fact that it is possible to apply this principle in written (computerized), group-administered tests, without intervention by a human assessor.

An essential precondition for applying the adaptive principle in computerized testing is that you must have a measuring tool or scale in which both the levels of skill of the candidates and the levels of difficulty of the items can be described. The following example shows how a common scale enables adaptive testing. Suppose you want to determine how high someone can jump. It seems obvious that you should first make a rough estimate of that person's capabilities in this regard. To do so, you use the rule of thumb that taller and slenderer people can jump higher than shorter and fatter ones, and that men in general can jump higher than women. Based on that information, you come to the initial conclusion that it is probably pointless to get a given person to try to jump over a bar that is lower than 60 cm or higher than 160 cm. After all, the result of an attempt with the bar at those heights is highly predictable and adds little or nothing to what you know about how well the person concerned can do the high jump. You decide to start somewhere in the middle of the range from 60 to 160 cm, for example at 110 cm. You observe that person closely during her or his attempt to jump the bar, and note that he clears it with ease. The first attempt has now given you a lot of information, i.e. that the person in question can probably jump a lot higher than 110 cm. You therefore decide to put the bar at 130 cm. The jump with the bar at this height fails, but only just. Your conclusion from this is that the person's capability will be closer to 130 cm than to the initial 110 cm and you therefore place the bar at 125 cm. If the person in question manages to jump this height, you conclude that his capability in the high jump is somewhere between 125 and 130 cm. This estimate is enough to keep you happy and you therefore end the session.

This jumping test is easy to translate into the educational situation. Instead of the skill to do the high jumping, a cognitive skill is measured such as e.g. arithmetic skills or language skills. The height of the bar is the level of difficulty of an item. Clearing the bar corresponds to a correct answer, whereas failing to clear it corresponds to giving a wrong answer. Then, just as in the examples above, in educational applications the successes and failures are used to make an estimate of the position of a person on the scale being used. However – unlike in the high jump – you cannot use a simple tape measure for assessing the cognitive skill. Determining the degree to which someone possesses a cognitive skill is done using a series of items that jointly form a scale. The more items in the series are correctly answered by a person, the higher his or her position on the scale and therefore the greater the ability level concerned. The problematic aspect of this is that the person's ability level and the difficulty of the items completed are inextricably intertwined to give the test score. Has someone answered a lot of items correctly or wrongly because he has a high or low ability level, or was it because the items were so simple or difficult? This problem can only be resolved by having the position on the scale determined not only by the number of correctly answered items, but also by the level of difficulty of those items.

Constructing the scale

Modern test theory allows us to resolve the problem stated above by applying models that give an explicit description of the relationship between the level of difficulty of an item and the ability level of a person. There are various models and we will describe one of them here. According to that model, the probability of a correct answer is exactly 50 percent if the ability level of a person is equal to the level of difficulty of an item (both being measured on the same scale). The example of the high-jump used above can help explain this. If we set the bar for the high jump exactly as high as the athlete can clear (i.e. at a height that is the average of all best heights that the person has cleared in various sessions), then we can expect that the athlete will knock the bar off in exactly half his or her attempts at this height and will clear it in the other half. If the level of difficulty of the item exceeds the ability level, then the probability of a correct answer becomes less than 50 per cent (the bar is knocked off more often than the athlete clears it). If the level of difficulty of the item is less than the ability level, then the probability of a correct response is greater than 50 per cent (the athlete clears the bar more often than it is knocked off). The relationship between the level of difficulty and the skill is described with a mathematical model that specifies the probability of a correct answer being given by a person with a given ability level. This probability is, of course, dependent on characteristics such as the degree of difficulty and perhaps other characteristics too, such as the discriminating power of an item. The model is given in the following equation (Hambleton & Jones, 1993):

$$p_i(\theta) = P(X_i = 1 \mid \theta) = \frac{\exp(a_i(\theta - \beta_i))}{1 + \exp(a_i(\theta - \beta_i))};$$

In this equation θ is the ability level of the person, β_i the level of difficulty of item i and a_i the discriminating power of item i.

With the help of data collected in pilot testing, the model can be examined to see if it provides a good description and prediction of the respondents' answers. The relative degree of difficulty is estimated for all items that meet the criterion. A scale is then obtained by ordering the items according to their difficulty. When a person sits a test that is put together from what is known as a 'scaled' item bank (i.e. a database in which the items behave consistently with the model described above), then the test result can be used to estimate the position of that person on the same scale as the one used to express the difficulty of the items.

How does CAT work?
Figure 1 gives a graphical representation of the process of administering an adaptive test. The progress of the test is shown on the horizontal axis. The vertical axis is used to display both the estimated ability level of the person (shown as circles) and the level of difficulty of the items (shown as crosses). The dotted line running parallel to the horizontal axis represents the performance standard, or the ability level that candidates must possess if the test result is to be positive. One weak point in the administration of an adaptive test is the start, because there is no information available at that point about the candidate's ability level and as a consequence the testing algorithm cannot select an item that gives the best match. Random selection of the first item is a widely-used solution, but there are other options too. The example in Figure 1, for instance, shows that the adaptive process only starts at the selection of the fourth item. The first three items are drawn at random from a subgroup of relatively simple items. This is a way of forcing the test to start with simple items, for example to reduce any anxiety about taking the test.

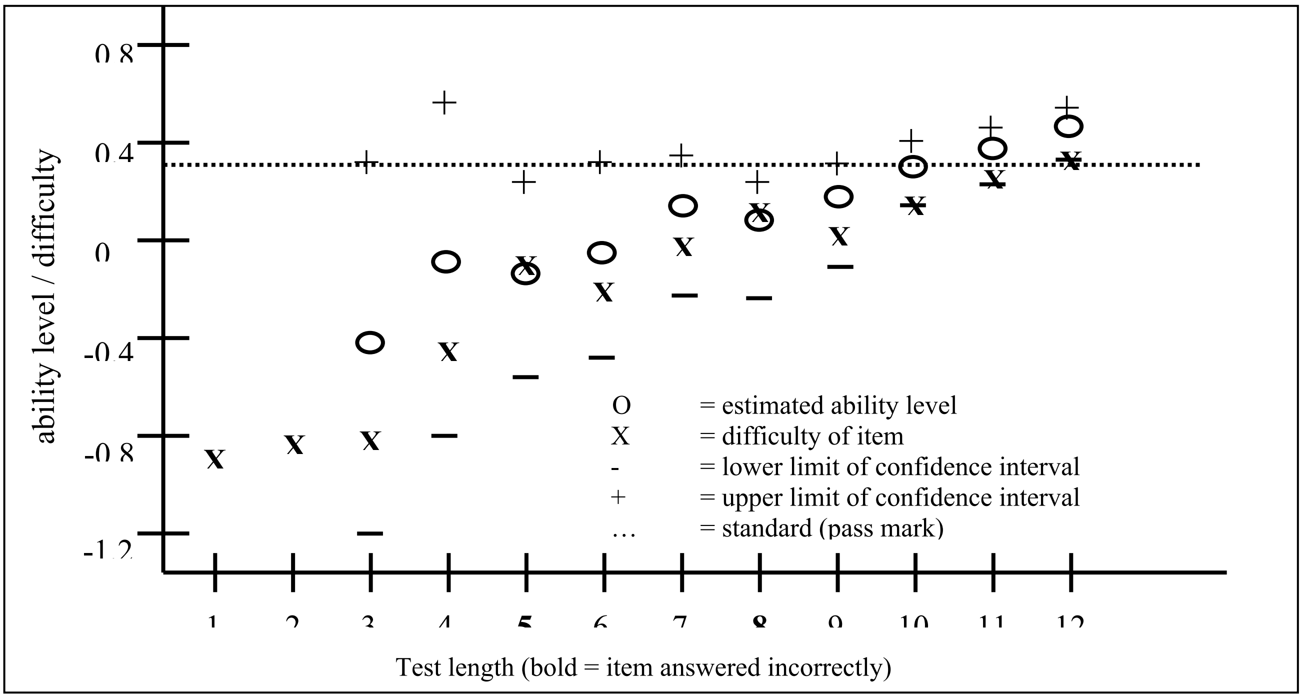


Figure 1: Progress of adaptive testing for a (fictitious) candidate

After the candidate has answered the third item, the first assessment of the ability level is made. This estimate cannot, of course, be particularly accurate after answering just three items, and the programme therefore estimates not only the ability level but also the uncertainty in the measurement, which it uses to define a confidence interval for the estimated ability level. The plus and minus signs represent the upper and lower limits of the confidence interval respectively. The confidence interval states – to a degree of certainty that you can select yourself – that the actual value for the candidate's ability is within the upper and lower limits of the interval. The actual ability level is that person's ability level on the scale used. It is easy to see from Figure 1 that the accuracy of the estimates of the ability level quickly becomes higher as the number of items answered increases.

There are various ways of terminating an adaptive test session. The simplest is, of course, that a fixed length is defined for the test. In the example in Figure 1, a dynamic termination rule has been adopted. That is to say, the test session ends as soon as the confidence interval around the most recent estimate of the skill is entirely above or entirely below the performance standard. In this case, that situation arises after the twelfth item has been answered. Now, it is possible to conclude with 90% (in this specific case) confidence that the true value for the candidate is above the cut-off mark used. It can be concluded with a high degree of certainty that the candidate does possess the skill concerned. The adaptive nature of the test process itself can be seen from the positions of the crosses and circles on the scale. The cross in any given column generally has a position on the scale that is somewhere close to that of the circle in the column to its left.

This special way of assembling tests has some attractive benefits. The key advantage is the greater efficiency of the testing process. Because the level of difficulty of each item administered is finely matched to the ability level of the candidate, the same accuracy of measurement can be achieved with fewer items than in a longer, conventional test. The literature often refers to reductions in test length of 40 percent or more (Wainer *et al.*, 2000).

Another benefit is that students are not confronted with tests that are much too difficult or much too easy. That is especially important in situations where the levels of skill to be measured are highly variable within the group. Conventionally constructed tests are too easy or too difficult for a large proportion of candidates in such situations. That not only leads to feelings of frustration or boredom, but also to imperfections in the measurement, because items that are far too difficult or far too easy do not add much information to what was already known before the item was answered. Now, every candidate gets a test that is challenging at his or her level. Another benefit that catches the imagination is that every candidate takes a different test. The risk of students passing information about test content on to one another is thereby considerably reduced, which paves the way for much more flexible test planning. Because test scores are always converted into estimated abilities on the ability scale that has been constructed, performance results on different tests can always be compared to one another. That is not only convenient when comparing the performance of different candidates, but also for assessing the progress of the individual candidates.

Description of the WISCAT-pabo testing package

Figure 2 shows a highly simplified diagrammatic representation of the testing package (leaving the technical infrastructure aside). WISCAT-pabo consists of four components that jointly handle the composition, actual testing, assessment and reporting of candidate testing. It works basically as follows. The testing algorithm is a computer programme that contains rules for the way in which the computer-controlled test session assembles a test from an item bank. In addition, the algorithm uses the answers supplied to give a continuous assessment of the skills of the candidate. The item bank contains the item texts, associated illustrative material and item characteristics (such as degree of difficulty) for each item. The module for administering the test presents the items on the screen one by one, scores the answers given by the candidate as right or wrong, and passes this result on each time to the testing algorithm. When the testing algorithm terminates the test, the reporting module handles feedback on the results to the candidate and the teacher. More detailed attention will now be paid to each of these components.

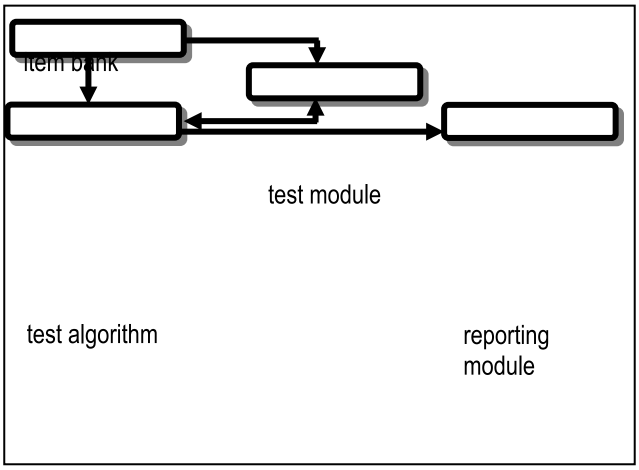


Figure 2: Schematic representation of WISCAT-pabo

The item bank

The item bank contains nearly 900 items that in total cover the functional aspects of the concept of 'arithmetic ability'. Table 1 shows which subject domains are represented in the item bank and how many items it contains.

Sub domains	**Number of items**	**Number that are mental arithmetic**
1. Basic operations such as addition, subtraction, division, etc.	202	140
2. Operations using fractions, percentages, ratios and decimals	341	127
3. Measuring using simple and composite units	141	15
4. Geometry. Interpretation of maps and spatial figures	90	
5. Ordering, representing, summarizing and interpreting statistical data	31	
6. Describing relationships as formulae (in words) and using these for calculation. Reading and interpreting graphs and tables.	53	
TOTAL	**858**	**282**

Table 1: Description of the item bank in WISCAT-pabo

WISCAT-pabo generates on-screen tests. This naturally imposes restrictions on the types of items that can be asked. When constructing the items, two types of items were used that are particularly suitable for computerized scoring: the multiple choice item and the short answer item. The latter type of item is one in which the candidate has to respond by entering a single number, word or symbol into a response field. It is clear that this type of item cannot be used to gain insights into the approaches used by the candidates to solve arithmetic items. Neither is it required: the purpose of WISCAT-pabo is to determine whether prerequisite knowledge and skills have been acquired by the students, rather than detecting any learning difficulties or misconceptions relating to arithmetic.

All the items have been tested in the target group. In total, about 2,500 first-year PABO students took part in what was referred to as 'pilot testing'. The data obtained in this pilot testing allowed to estimate the degree of difficulty and the discriminating power of each item, followed by a check that the model chosen (see the section on 'Constructing the scale') gave a good description and prediction of the pilot test results. Items that did not 'behave according to the model' were then removed from the item bank. The remaining items could then be ranked according to difficulty in order to create a scale for measuring arithmetic skills. What the scale concept means in concrete terms in this case is that a student who answers a particular arithmetic item correctly is more likely to answer items with lower scale values also correctly. However, the more arithmetic items with higher scale values are presented, the more the probability of a correct response diminishes.

Constructing a scale of arithmetic ability only makes sense if a point on that scale can be indicated that has to be achieved to pass the test; in this particular case a critical score that can be used as a basis for continuation of the course of study. That point was determined using a standard described in qualitative terms by an expert committee of PABO arithmetic teachers: "First-year PABO students must be able to do arithmetic by the end of their first year of study as well as a good pupil from the final year at primary school." The expert committee took a 'good pupil' to be one whose arithmetic abilities are in the top 20 percent of final-year primary school pupils. By also having part of the item bank answered by a representative random sample of final-year primary school pupils, the distribution of ability in that population could also be shown on the scale constructed for the PABO students. In the distribution of ability, a point was located on the scale that exceeded the capabilities of 80 percent of the final-year primary school pupils. That point (value 103 on a scale ranging from 0 to 200) is the nationwide standard (pass mark) that WISCAT-pabo uses for taking pass/fail decisions.

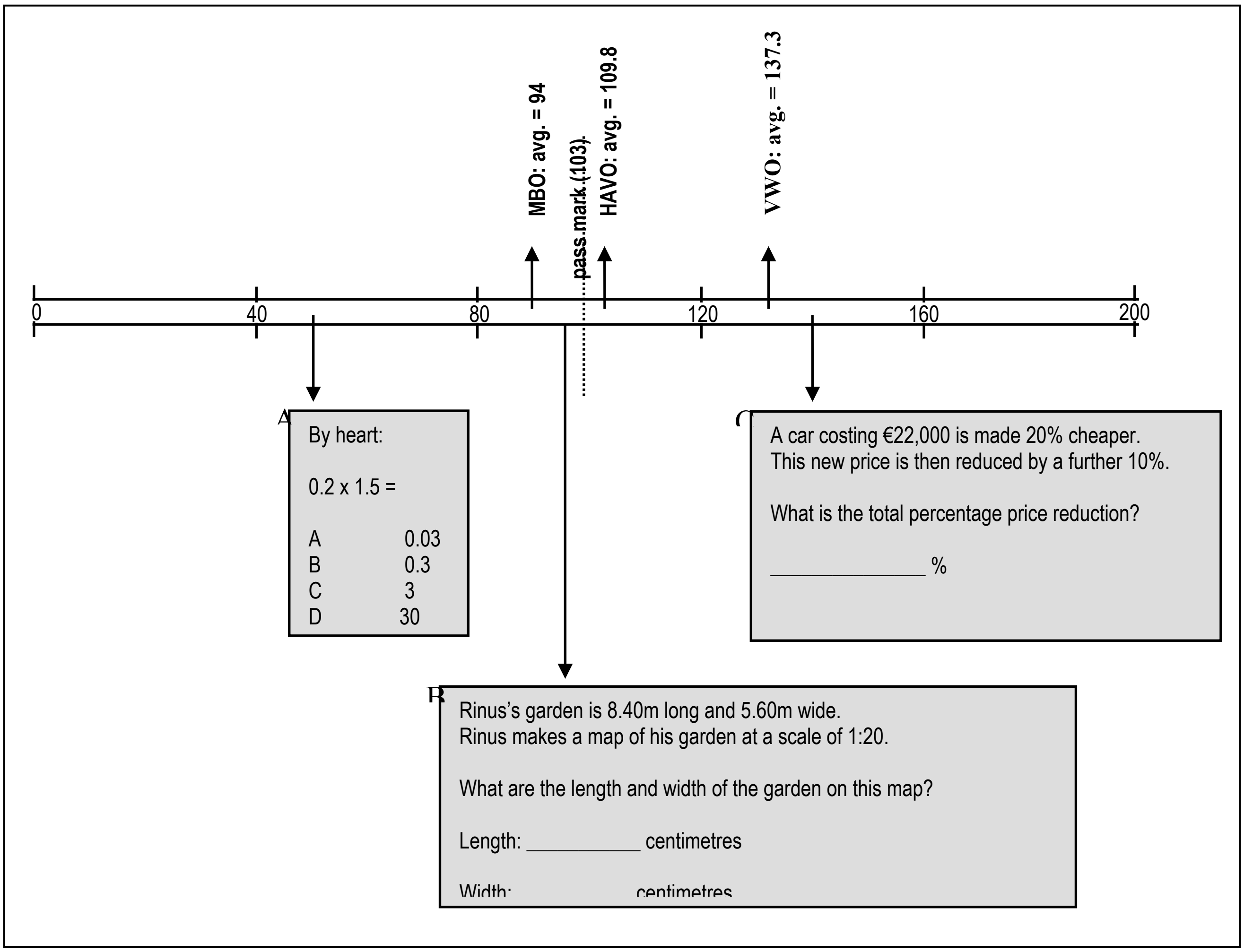
Figure 3 gives a highly simplified graphical representation of the arithmetic ability scale that has been constructed. The following points are marked on the scale:

- the average value, as determined during the first operational year of the test, for first-year PABO students, broken down according to previous education
- the national standard (cut-off mark) used in the WISCAT-pabo testing package
- the level of difficulty of three example items

From this data, the underlying model can be used to calculate the probability of a correct answer for each of the three items shown. For the average candidate with prior education at the MBO (vocational secondary), HAVO (general secondary) and VWO (general/pre-university secondary) levels, the probabilities are as follows:

Item A: 0.76 (MBO); 0.84 (HAVO); 0.93 (VWO)
Item B: 0.44 (MBO); 0.57 (HAVO); 0.76 (VWO)
Item C: 0.16 (MBO); 0.25 (HAVO); 0.44 (VWO)

Once a candidate's ability level has been estimated on the scale, the same method can be used to calculate the probability of a correct response to any item in the item bank.



By heart:

0.2 x 1.5 =

A 0.03
B 0.3
C 3
D 30

A car costing €22,000 is made 20% cheaper.
This new price is then reduced by a further 10%.

What is the total percentage price reduction?

_____________ %

Rinus's garden is 8.40m long and 5.60m wide.
Rinus makes a map of his garden at a scale of 1:20.

What are the length and width of the garden on this map?

Length: __________ centimetres

Width: centimetres

Figure 3: Simplified graphical representation of the arithmetic ability scale

The testing algorithm

The test consists of two parts. In the first part, 15 mental arithmetic items are given for which the candidate has precisely 15 minutes. How the available time is distributed across the items is left to the candidate. After a quarter of an hour, this part of the test is automatically closed down. Items that are not answered by then are scored as incorrect. The second part of the test consists of 35 items. While working on these items, candidates are allowed to do their working out on paper. A calculator may be used in some cases. Where this is allowed, a calculator appears on the screen.

The testing algorithm uses an adaptive method to put together a test from the item bank. This means that each subsequent item is always chosen so that it gives the optimum fit for the most recent skill level estimate. This does have certain drawbacks, along with the benefits. One disadvantage is that you have no control over the content of the test that is being put together. That turns out to be incompatible with the desire to be able to report on the candidates' performance using a profile of scores (separate reporting about important subdomains of arithmetic). That is because this requires particular subject matter to be represented in the test by a sufficient number of items. In order to achieve this, restrictions are imposed on the

testing algorithm to ensure that the selection of items does not look only at the resulting measurement accuracy, but also at the coverage of the test grid.
The items must therefore be selected in such a way that sufficient items are offered from each part of the subject matter for which a profile score has to be given: mental arithmetic; basic skills; fractions, percentages, ratios and decimals; and measurement and geometry. The restrictions imposed do mean that the adaptive algorithm will not always select the item that would be the best one from a purely psychometric point of view.

The functioning of the testing algorithm was checked extensively in the pre-operational phase in what are known as simulation studies. In a simulation study, the model being employed (see the equation above) is used to generate item responses from very large numbers of fictitious candidates. This provides an easy way of investigating whether or not the testing algorithm functions as the designers intended, for example whether the tests that are generated are indeed structured according to the test grid specifications or e.g. how accurate the decisions taken about the candidates are, when based on the test result.

When taking decisions about candidates based on the test results, two types of errors can be made:

- The first type of error occurs (in this concrete case) if the measured (estimated) ability level of the student meets the nationally defined standard while the actual true value does not. The student has then passed, inappropriately.
- The second error is when the measured ability of the student does not meet the nationally defined standard, while the actual value does. In such cases, the student is unfairly rejected.

Because the true value can never be known, it is difficult to evaluate the quality of the pass/fail decisions. However, simulation studies can offer a way out. Based on the model chosen, it is possible to simulate taking the test, with both the true skill level (the value that the researcher chooses and that the simulation starts at) and the estimated ability level (the value estimated after the last test item has been 'answered') being known, so that they can be compared with one another.. If this is repeated a large number of times for 'candidates' of varying ability levels, a good picture can be obtained of the theoretically achievable quality of the decisions.

The results of these simulation studies were used to produce a table for the accuracy of the decisions made (see Table 2). This is applicable for all tests that are taken using WISCAT-pabo. The table gives the percentages of correct and incorrect decisions based on test scores that were generated with the test model employed.

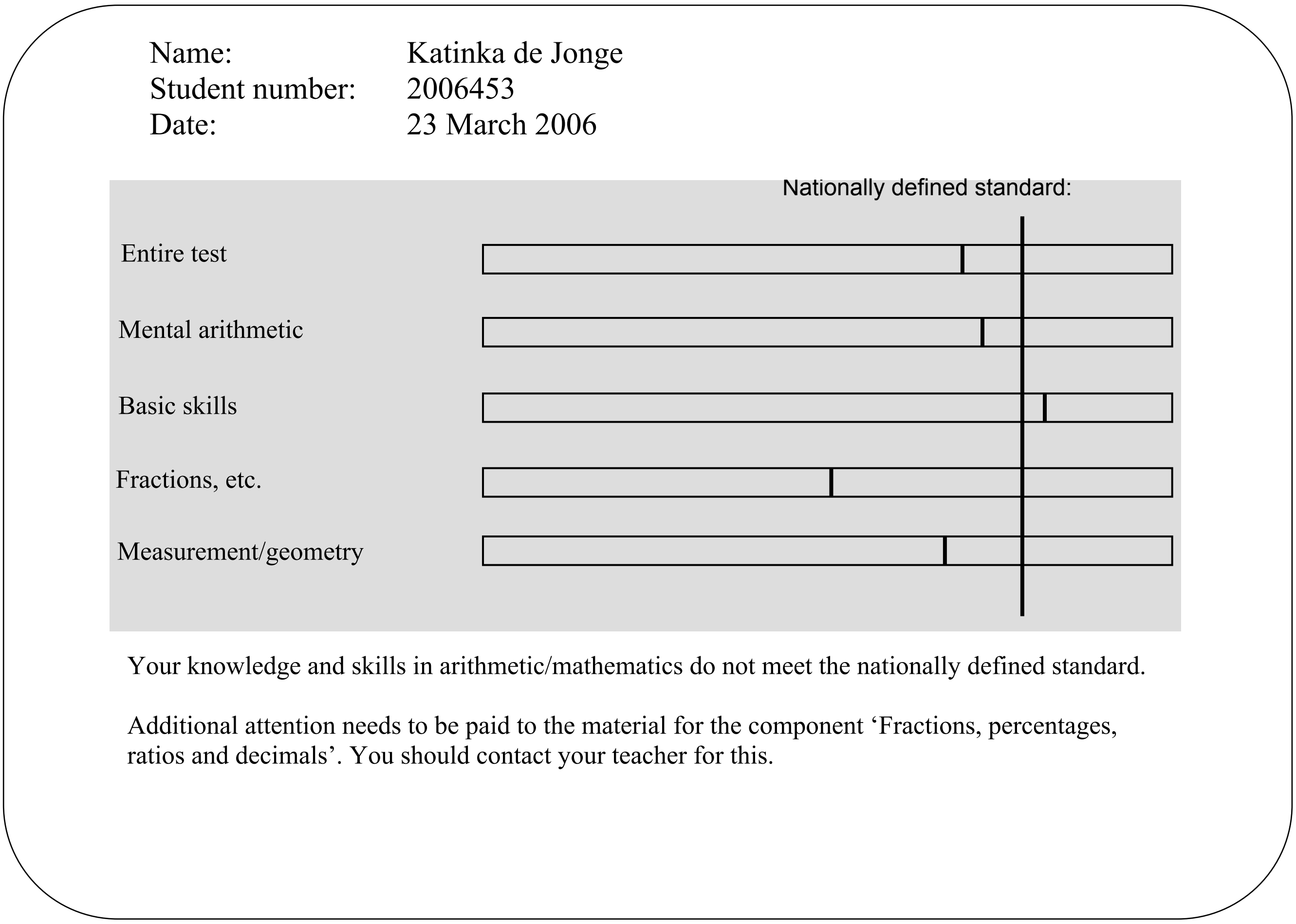
		estimated ability level	
		fail	pass
actual ability level	fail	50.6	4.6
	pass	4.2	40.6

Table 2: Percentage of correct and incorrect decisions

The table shows that a correct decision is made in over 90 percent of cases and that the two types of incorrect decisions occur about equally often. As stated, these results show the *theoretical* accuracy of the decision-making. In reality, people will not always behave conform the test model used here. The percentages are therefore no more than an indication for the proportion of erroneous decisions that can be expected in real application situations.

The reporting module
As soon as the final test item has been answered, the candidate is shown the result on the screen. Students can easily read off from it whether their knowledge and skill meet the nationally defined standard. This is decided by the result on the whole test. The results for the subdomains are also reported. If the result for one subdomain is significantly lower than the whole result, then the programme states this explicitly.

In Figure 4 an example of a report is given:



Name: Katinka de Jonge
Student number: 2006453
Date: 23 March 2006

Your knowledge and skills in arithmetic/mathematics do not meet the nationally defined standard.

Additional attention needs to be paid to the material for the component 'Fractions, percentages, ratios and decimals'. You should contact your teacher for this.

Figure 4: Report for the Student

For the teachers or other supervisors, the reporting module produces more numerically oriented results that can be formatted as preferred, either individually or for each group (see Figure 5).

7

Results for Katinka de Jonge **Student no.: 2006453**			
Nationally defined standard: 103			
Attempt	1	2	3
date	06-09-05	12-01-06	06-06-06
Test result	**88 (F)**	**97 (F)**	**105 (P)**
Mental arithmetic	96	103	109
Basic skills	95	100	111
Fractions, etc.	84	99	102
Measurement/geometry	69 *	72 *	99

Figure 5: Report for teacher. Individual summary of arithmetic ability. (F = fail, P = pass, * = significantly lower than the candidate's overall test result)

First results

During the 2006-2007 academic year, the test was taken one or more times by 10,978 PABO students in their first year. The test was administered a total of 17,610 times. Candidates were permitted to take the test up to a maximum of three times in order to achieve the national standard. A small number of students even took the test four or more times: because some PABO schools impose higher demands than the national standard, there were cases where candidates who already met the national standard still took additional tests.

Previous education	Number of tests taken	Average	Standard deviation
MBO (secondary vocational)	7817	94.0	27.7
HAVO (general secondary)	8172	109.8	26.4
VWO (secondary/pre-university)	1453	137.3	32.6
Unknown	168	102.4	32.9
Total	17610	105.0	30.2

Table 3: WISCAT-pabo and previous education

Table 3 gives the mean values and standard deviations of the WISCAT-pabo scores as a function of the students' previous education. The differences in the average scores between candidates with different prior education are statistically significant and in the direction that would be expected. Previous education level has a large effect on the WISCAT-pabo scores.

Of the almost 11,000 candidates, 75.6% finally achieved the national standard. The results for each time the test was administered are given in Table 4. The numbers and percentages of success are also given for each type of previous education.

Attempt	Number of candidates passing	Percentage passes
First	5740	52.3
Second	7358	67.0
Third	8237	75.0
All attempts, by previous education level:	8299	75.6
MBO (secondary vocational)	2614	60.5
HAVO (general secondary)	4390	83.5
VWO (secondary/pre-university)	1207	94.9
Unknown	88	68.8

Table 4: Passing percentages WISCAT-pabo by previous education type

After the first attempt, only a little over half the candidates have passed. After resits have been taken, this percentage finally increases to three quarters of all students in the first year teacher training college. There are big differences in the percentages of passing depending on previous education. Almost all former VWO pupils, over 80 percent of former HAVO pupils and 60 percent of former MBO pupils meet the defined standard.

First experiences of the students

The experiences of the students have not yet been studied systematically. The items listed below are the experiences of students who expressed their dissatisfaction with WISCAT-pabo on their own initiative. It is not yet known whether their problems are typical of the entire group.

Stress due to "tailored testing": The key argument in favour of adaptive testing is the greater efficiency with respect to traditional testing. This makes it possible to use shorter tests than normal, while retaining the same accuracy of measurement. This is achieved by offering "tailored testing". Test designers generally assume that it is obvious that students prefer it when a test consists of items that are neither too difficult nor too easy. It was therefore surprising that some of the students reported that they became nervous when they believed that they were able to tell from the degree of difficulty of the items that they had given a lot of wrong answers. In fact, the percentage of incorrect answers that would be anticipated when sitting adaptive tests is fairly constant, at about 50 percent. The solution for this must provisionally be sought in better explanation for the student about the working method and the background of adaptive testing.

Restricted freedom: Stress and anxiety about the test were also reported, predominantly as a consequence of the severely restricted freedom for the student when sitting the test. Because the processes of taking an adaptive test and composing it are simultaneous, there is no (electronic) test booklet available during the test that the students can look at and use to determine the order in which to tackle the items for themselves. There is no solution for this as yet: adaptive testing is incompatible with freedom for the candidates to determine for themselves the order in which the items in a test are tackled. Explanation beforehand and practice also seem to be the best remedies here.

Focused on the results alone: Some students have remarked that WISCAT-pabo is excessively oriented towards the results of arithmetical processes, whereas for (future) teachers it is precisely the focus on the processes themselves that is so crucial. It is true that WISCAT-pabo does not look at the arithmetical processes and that only the result counts. There are two reasons for this. First, it is as yet not possible to analyse and assess the candidates' arithmetical working out in a computer-controlled test. Secondly, given the objective of the test, this is not necessary. It is not about detailed diagnosis (detecting and categorizing misconceptions and learning difficulties); it is about an efficient way of distinguishing the students who do or do not have the minimum initial level of knowledge and skills required.

Finally

In this article, we have shown that complex measurement problems in the field of education can be beaten by combining two powerful techniques: modern testing theory (IRT) and computer-controlled testing. Item response theory has made it possible to develop scales on which both the ability level of a person and the level of difficulty of the test items can be expressed. The principle of the common scale made it possible to have tailored testing in examinations. The development of the personal computer allowed this to be done in practice, with the huge processing speed of a computer making it possible to handle composing a test, administering the test and scoring the test as virtually simultaneous processes.

CAT is no longer in its infancy and interest is increasing worldwide in the use of this powerful technique for concrete testing purposes. The advantages are often highly attractive, as was the case for testing arithmetic skills in first-year student teachers. The specific measurement problems associated with this were beatable, thanks to the use of CAT. The first experiences are positive, but also seem to teach us that acceptance of these kinds of complex testing concepts by those involved is not a *fait accompli*. It is important that sufficient time and resources are made available so that the working methods, benefits and idiosyncrasies of CAT can be explained to all those involved.

References

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: issues and practice, 12*(3), 535-556.

Straetmans, G.J.J.M. & Eggen, T.J.H.M. (2005). Afrekenen op rekenen: Over de rekenvaardigheid van pabo-studenten en de toetsing daarvan. *Tijdschrift voor hoger Onderwijs, 23*, 3, 123-139.

Suen, H.K. (1990). *Principles of Test Theories*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. (Ed.), (2000). *Computerized Adaptive Testing: A primer* (Second edition.) Hilsdale, NJ: Lawrence Erlbaum Associates.

The authors:

Theo Eggen & Gerard Straetmans
Cito
Nieuwe Oeverstraat 50
P.O. Box 1034
6801 MG Arnhem
The Netherlands
T +31 26 352 11 11
F +31 26 352 13 56

E-Mail: *Theo.Eggen@cito.nl*
WWW: *www.cito.nl*

Dr. Theo J.H.M. Eggen is measurement statistician. He works as Senior Research Scientist at the Measurement and Research Department of Cito in the Netherlands. He is the project director of the Cito research project on computerized adaptive testing (CAT). He is professor of Psychometrics at the University of Twente in The Netherlands.

Dr. Gerard J.J.M. Straetmans is a senior assessment specialist at Cito, Arnhem (The Netherlands). He is also professor of educational assessment at Saxion University of applied sciences.

Issues in Computerized Ability Measurement: Getting out of the Jingle and Jangle Jungle

Oliver Wilhelm
Humboldt University Berlin

Abstract

Computerized ability measurement is not fundamentally distinct from other forms of educational and psychological measurement. The core issues frequently have to do with the constructs under investigation and the theoretical and empirical arguments that convince us that a certain test is measuring a certain disposition of a person. The issues we encounter in trying to provide support for such arguments are often not very well developed in the literature. In this contribution I will focus on two neglected validity fallacies that both provoke the development of substantive ideas and claims about why a measurement device is measuring a certain disposition. Turning our attention to these fallacies more frequently and more profoundly is likely to a) help in overcoming the serious problems associated with committing these fallacies and b) facilitate dialogue about what a test really measures.

The measurement of psychological attributes in general and the assessment of human cognitive abilities in particular is done almost ubiquitously to go beyond the observed data in an attempt to make general inferences about dispositions. In most cases cognitive ability measures are expected to inform us about the intensity of some trait underlying observed performance (Borsboom, 2008; Borsboom, Mellenbergh, Van Heerden, 2003). Usually we conceive this trait to be causal for observed performance (Borsboom, Mellenbergh, Van Heerden, 2004). Therefore, it is common to think of our measures as being instances of reflexive measurement models (Bollen, 1989). In the recent past, it has become more prevalent to use methods that are suited to appropriately transfer such a conception of abilities into adequate measurement models.

However, we are convinced that an adequate understanding of our measures also relies on developing, testing, and failing to reject sound and theoretically motivated ideas about the relations between constructs. One reason why this is so important has to do with the excess meaning assigned to constructs relative to what we do on the assessment level. The dispositions we want to measure have labels like “General Cognitive Ability”, “Reasoning”, “Student Achievement in Mathematics in 9th grade”, “the ability to solve problems and think critically in a digital environment”, “the ability to use and understand English as it is spoken, written, and heard in college and university settings” and so on. At the beginning of this list we used highly general terms – provoking the questions a) what are the specific measures used for assessment, b) is the assessed domain well defined, and c) will alternative, equally well suited measures deliver the same inferences about the intensity of traits, differences between persons etc.. Towards the end of this list we were using more specific descriptions, leaving less but still many degrees of freedom on what to use as assessment devices. However, more specific descriptions of what a test supposedly measures provoke questions like a) how does this ability relate to abilities with similar labels, b) how is this ability distinct from abilities with dissimilar labels, and c) how can this ability be embedded into a nomological net of individual differences constructs.

In this contribution we want to focus on two prominent but neglected issues in this interpretation process of ability measurement. These two issues have been labelled *Jingle Fallacy* (Thorndike, 1904) and *Jangle Fallacy* (Kelley, 1927). The Jingle Fallacy refers to situations in which instances are treated as identical although the only known communalities of those instances are their similar labels. The Jangle Fallacy refers to situations in which instances are treated as different although the only known distinctions between these instances are their diverging labels. Obviously, a Jangle Fallacy is also present if prior work in a (related) field is ignored and a supposedly new construct is declared new because of a new construct label.

In the context of educational and psychological measurement these two fallacies refer to the distinction between essentially deliberately chosen construct labels and theoretically and empirically supported construct meanings. Over the last decades, both fallacies have not been treated very prominently in research on educational and psychological measurement. Some exceptions (amongst them Block, 1995; Heyman & Dweck, 1992; Marsh, 1994; Wilhelm, Schulze, Schmiedek, & Süß, 2003) illustrate the generality of the problem. Committing a Jingle or a Jangle Fallacy is a serious flaw in educational and psychological measurement: If you commit a Jangle Fallacy your work is akin to the reinvention of the wheel. Either a lack of knowledge or of concern about the work of other scientists or a lack of required data demonstrating divergent validity between new and established measures/ constructs is the cause of the fallacy. If you commit a Jingle Fallacy you transfer available evidence to a new measure although in reality what you conceive as being an instance of an established construct somehow diverges.

With respect to computerized measurement the approaches used in the assessment literature can be roughly classified as attempts to use the computer as a new test medium for existing constructs or as attempts to use the computer to test for new constructs. Using the computer as a new test medium suggests that the only difference in the measurement process comes from a supposedly construct irrelevant aspect. Using the computer to test for new constructs presumes that technology allows you to assess a disposition or trait that could not be measured adequately before or without the new technology. In both cases it is crucial to address Jingle and Jangle pitfalls in research that are expected to support messages about the quality of the measures. In our opinion, this is done to an insufficient degree in ongoing research on computerized ability measurement.

Jangle Fallacy

Let us first come to some examples of the Jangle Fallacy. The Jangle Fallacy refers to situations in which two constructs with different labels are actually the same. The fallacy refers to the problem that already established constructs are "reinvented" or rediscovered. The basic logic in establishing a new construct is that it is crucial to convincingly show that the new construct can be measured appropriately and that the individual differences captured with a measure are somehow new. The first part of this challenge in establishing new constructs can be accomplished through showing that a meaningful and relatively parsimonious measurement model can account for individual differences data of a sufficiently large sample from the application population. The second part deals with novelty. We think there are at least three stages of novelty that a measure can meet and we will go through these stages sequentially and provide examples.

The first stage of novelty tries to show that what is captured with a new measurement device is different from apparently competing measurement approaches. To us it is obvious that in the self-report area - for instance - in the realm of openness for new ideas (Costa & McCrae, 1992), need for cognition (Cacioppo, Petty, Feinstein & Jarvis, 1996), rational experiences (Epstein, Pacini, Denes-Raj & Heier, 1996), typical intellectual engagement (Ackerman, 1994; Goff & Ackerman, 1992, Wilhelm, et al., 2003), understanding (Stumpf, Angleitner, Wieck, Jackson & Beloch-Till, 1985) there are serious issues in distinguishing these self-report dimensions from each other on the construct level (see also, Ackerman, & Gogh, 1994; Rocklin, 1994; Saucier, 1992; Trapnell, 1994). This issue becomes salient if you try to a) assign construct labels to construct descriptions or b) construct descriptions to sample items of any of the above mentioned instruments. In other words: Degrees of endorsement to an item like "I enjoy philosophical discussions" might be taken as indicative for several constructs. Relations between scales of such items reflect reliability or item sampling bias rather than construct validity. If we cannot meaningfully make discriminations between items belonging to distinct constructs then an empirical analysis showing that correlations between latent variables of such constructs are visibly below unity is obsolete. If that is the case we would be at a loss because it is then likely, that the item compilation does assess more than a single dimension. Therefore, with respect to self-reports of cognitive motivation the first stage of novelty-testing could not be counted as completed.

The second stage of novelty testing is more challenging. Here an attempt is made to show that a supposedly new construct is not only different from a single established construct but also from combinations of various established constructs. In other words, the new construct is not only distinct from any single established construct but also from any combination of established constructs. For example in the area of ability measurement the effects of time restrictions imposed on standard measures of maximal behaviour have been a controversial topic for a long time (Furneaux, 1960; Odoroff, 1935; Paterson & Tinker, 1930; Peak, & Boring, 1926; Thurstone, 1937). Some time restrictions on the completion of ability measures have to be applied for pragmatic reasons. These restrictions can have varying degrees of strictness and contingent on the strictness means and covariance of observed variables might change. If time constraints keep subjects from completing an item they have no opportunity to improve their score. On traditional number-correct scores they will therefore perform worse then they would if there were no time constraints. More importantly time constraints might affect the covariances of a measure. It is possible that time constraints make clerical or mental speed more relevant for the scores obtained in complex ability measures. We tested this hypothesis in a larger study on individual differences (Wilhelm, & Schulze, 2002) with a model that treated time-constrained reasoning ability as a linear function of clerical / mental speed and untimed reasoning ability. The residual of the factor "time-constrained reasoning" was fixed to zero. This model provided a decent account of the data and was not worse than a model allowing the time-constrained reasoning factor to have a residual variance after being regressed on clerical speed and untimed reasoning ability. One of the two predictors was clearly insufficient to account for the criterion. To sum up, we showed that a linear function of two established abilities is sufficient to account for a potential distinct ability. Therefore, the first but not the second step of novelty testing was completed by timed reasoning ability.

To give an example that is more computer assessment specific: How about complex problem solving? Is there any evidence that complex problem solving completes the second step of novelty testing? We are convinced that complex problem solving can be viewed as a linear function of reasoning ability and relevant knowledge in a given scenario. It is up to the proponents of new constructs to show that they clear that hurdle of novelty testing we describe here by collecting appropriate data and by running conclusive analyses.

At the third stage of novelty testing an attempt is made to show a) that a supposedly new construct cannot be regressed without rest on combinations of established constructs and b) that the new construct has some incremental validity in the prediction of relevant outcomes (Sechrest, 1963). Incremental validity refers to the desideratum that a supposedly new construct improves on diagnostic decisions. A convenient and convincing way to show incremental validity of "new" constructs is by showing improved prediction of latent variables of criteria. The improvement is relative to a model that excludes the new construct. *Situational judgment tests* (SJTs) might serve as an example here. SJTs are simulations requiring a participant to assess, evaluate, and judge how to respond to a hypothetical problem (Motowidlo, & Tippins, 1993) that usually occurs in a work context. SJTs are frequently used in employment contexts in the hope that they provide incremental validity over meta-analytically firmly established predictors like general ability measures, job knowledge, and conscientiousness in predicting prospective success on the job or in training. Taken together, the evidence suggests that SJTs succeed in the third stage of novelty testing (Clevenger, Pereira, Wiechmann, Schmitt, & Harvey-Schmidt, 2001; McDaniel, & Nguyen, 2001; O'Connell, Hartman, McDaniel, Grubb, & Lawrence, 2007). Obviously, completing stage three doesn't imply that there are no problems with the validity of SJTs. For example, one prominent issue is the challenge to substantiate what the communality of a variety of SJTs actually reflects and how to interpret such a construct.

Establishing a new construct is everything else than a small endeavour. Therefore, it is no surprise that convincingly completing stage two or three of novelty testing (as described above) is not happening very often. Generally, we would argue that if a latent variable is just a linear function of one or more established constructs there is no point in maintaining such a construct. A deviation from such a strict policy might be indicated if the "new" construct allows for superior measurement. For example, this criterion would be met if test items indicative of the "new" construct can be a) generated automatically, b) administrated adaptively, or c) return more reliability per test time etc.. Beyond such methodological and pragmatic concerns an informed preference between two essentially perfectly correlated constructs might also be legitimated on substantive grounds. Fluid intelligence and working memory might be such a couple (Ackerman, Beier, & Boyle, 2005; Kyllonen, & Christal, 1990; Oberauer, Schulze, Wilhelm, & Süß, 2005). Despite their (close to) perfect correlation working memory might be preferred over fluid intelligence because it might be a theoretically more coherently developed construct.

Obviously, policy reasons might come into play, too, if a construct isn't really new. Imagine that latent variables derived from two tests A and B are perfectly correlated with each other. Now also imagine that test A has instantiations of item content that appear face valid in some context of educational assessment – test A might be a student achievement test, for example. Please imagine that test B on the other side might be a number series test from a standard intelligence test lacking the face validity test A possesses. It is possible that latent variables for both measures are perfectly correlated but we might be tempted to accept test A as the appropriate measure of student achievement and would declare that test B is a quantitative reasoning test and inappropriate for that purpose. The perfect correlation between both latent variables tells us that we committed a Jangle Fallacy. Nevertheless, we would deem it acceptable to prefer test A for a pragmatic testing purpose over test B as long as it is acknowledged that both tests seemingly measure the same underlying ability.

We think it is very important to integrate research findings into a nomological net of established individual differences constructs. In many cases contemporary research has the potential to widen and deepen our knowledge of the personality traits at stake. However, we are afraid that addressing issues of Jangle fallacies by integrating approaches from a variety of fields is not appropriately rewarded by the scientific community. Therefore, the Jangle fallacy is more frequently an obstacle to the quality of research than it should be.

Jingle Fallacy

Let us now turn to the Jingle Fallacy. The Jingle Fallacy refers to the situation that two constructs with the same label are actually distinct. The fallacy refers to the problem that two superficially similar constructs are not assessing the same underlying trait. In this situation a serious threat is imposed on the validity of both constructs. In terms of classical validity research the correlation between two superficially identical traits is expected to be on the level of reliability estimates or, put more adequately, latent variables for both traits should be correlated at unity. If this is not the case either one or both instantiations of the supposedly identical traits are flawed. Such a situation might provoke more profound thinking about the construct at stake, for example, in multi-trait multi-method studies.

One prominent example of a Jingle Fallacy comes from computerized ability measurement and concerns the construct of mental speed. Mental speed can be defined as rapid scanning and responding in intellectually simple tasks (Horn & Noll, 1994) or as performance in so-called elementary cognitive tasks. A variety of labels is used in the literature to refer to such tasks – amongst them "clerical speed", "information processing speed", "mental speed", "decision time speed", "cognitive speediness" and the like (Danthiir, Roberts, Schulze, & Wilhelm, 2005). We will use the term mental speed as standing vicariously for all these labels. Importantly, empirical research shows that mental speed ought to be considered to be organized within a higher order factor structure model (Danthiir, Wilhelm, Schulze, & Roberts, 2005).

A purely psychometric definition of mental speed tasks is that it includes any cognitive task measured with items that are so easy that any person from the application population solves almost all of them correctly. There are two prominent scoring procedures for such tasks. The first scoring procedure times each response individually and expresses performance in a latency-per-correct-response metric. The second scoring procedure counts the number of correct response in a testing unit that implements a severe time-constraint. Performance is then usually expressed in a correct-response-per-time metric. Both metrics can be transformed into each other by inverting scores. In computerized measurement responses are usually timed individually – in traditional paper-and-pencil measurement items are usually presented listwise and a count of correct responses is applied.

Given that scores from both scoring procedures can be transformed into each other the distinction between various task classes might seem to be trivial. Indeed, when referring to mental speed constructs rarely amendments like "as assessed by administration technology Y" are reported in the literature. Therefore, the test medium can be expected to be irrelevant in the assessment of mental speed. Nevertheless, it is well documented that computerized and paper-based mental speed measures show rather small correlations even in circumstances in which equivalence across test media was an overarching goal (Mead & Drasgow, 1993). In this meta-analysis correlations between computerized and traditional paper-and-pencil measures of mental speed were on average r=.72 after being corrected for measurement error. This correlation is not much higher than what you can find, for example, between latent factors for working memory and mental speed in some samples. This implies that when it comes to mental speed the test medium is a decisive component. If we think about the test medium as a method factor and about different constructs as trait factors this worst case scenario can be summarized as the fact that monotrait-heteromethod correlations for mental speed are not higher than some of the heterotrait-monomethod correlations.

There might be a variety of reasons for this divergence and the disturbing results might be a good starting point to think more thoroughly about our constructs. For example, one might argue that mental speed should not be contrasted to other ability constructs it should rather be considered to be the second component to express mental work – the other component being mental accuracy. One might also argue that there is a variety of differences between both test media that are causal for fractions of the difference. For example, inter-trial speed, phasic alertness, item- vs. listwise presentation of items, difference of relevant psychomotor abilities and so on all might contribute to the observed divergence.

Discussion

In this contribution we attempted to highlight some issues in the interpretation of measurements we routinely perform. Two not very prominent validity fallacies might be "misunderestimated" in their relevance for ongoing psychometric research. We tried to argue that these fallacies deserve more attention than they currently receive.

The questions of validity we were addressing in this contribution are highly relevant for issues in computerized testing. It is important to realize that ensuring equivalence across test media is insufficient to show that a computerized measure has a property we label "construct validity". This is the case because most measures we use in ability measurement have a somehow dubious status with respect to their construct validity. Take the SAT for example. This measure is used a few million times every year and is unprecedented in terms of psychometric care that goes into the construction and maintenance of the test. Nevertheless, it is open to interpretation what the test really assesses (Bridgeman, 2005; Frey & Detterman, 2004). Efforts to establish a higher-order model that covers a substantial area of human behaviour (Carroll, 1993) are mostly data-driven and also they are indicative of a major effort they remain vague and imprecise in many ways. The last few decades have shown decent progress in the methods used to study human cognitive abilities. We hope that the future shows a similar progress on a substantial level.

References

Ackerman, P. L. (1994). Intelligence, attention, and learning: Maximal and typical performance. In D. K. Detterman (Eds.), Current topics in human intelligence. Vol. 4: Theories of intelligence (pp. 1-27). Norwood, NJ: Ablex.

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? Psychological Bulletin, 131, 30–60.

Ackerman, P. L. & Goff, M. (1994). Typical intellectual engagement and personality: Reply to Rocklin (1994). Journal of Educational Psychology, 86, 150-153.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. Psychological Bulletin, 117, 187-215.

Bollen, K. A. (1989). Structural equations with latent variables. NY: John Wiley & Sons.

Borsboom, D. (2008). Latent variable theory. Measurement, 6, 25-53.

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2003). The theoretical status of latent variables. Psychological Review, 110, 203-219.

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. Psychological Review, 111, 1061-1071.

Bridgeman, B. (2005) Unbelievable Results When Predicting IQ From SAT Scores. A Comment on Frey and Detterman (2004). Psychological Science 16, 745–746

Cacioppo, J. T., Petty, R. E., Feinstein, J. A. & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. Psychological Bulletin, 119, 197-253.

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. Cambridge, MA: Cambridge University Press.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey-Schmidt, V. (2001). Incremental validity of situational judgment tests. Journal of Applied Psychology, 86, 410-417.

Costa, P. T. Jr. & McCrae, R. R. (1992). Revised NEO personality and five factor inventory professional manual. Odessa, FL: Psychological Assessment Resources.

Danthiir, V., Roberts, R. D., Schulze, R., & Wilhelm, O. (2005). Approaches to mental speed. In O. Wilhelm, & R. W. Engle (Eds.), Understanding and measuring intelligence (pp. 27-46). London: Sage.

Danthiir, V., Wilhelm, O., Schulze, R., & Roberts, R. D. (2005). Factor structure and validity of paper-and-pencil measures of mental speed: Evidence for a higher-order model?. Intelligence, 33, 491-514.

Epstein, S., Pacini, R., Denes-Raj, V. & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. Journal of Personality and Social Psychology, 71, 390-405.

Frey, M. C., & Detterman, D.K. (2004). Scholastic assessment or g? The relationship between the Scholastic Assessment Test and general cognitive ability. Psychological Science, 15, 373–378.

Furneaux, W. D. (1960). Intellectual abilities and problem solving. In H. - J. Eysenck (Ed.) Handbook of abnormal psychology (pp. 67–192). London: Pitman Medical.

Goff, M. & Ackerman, P. L. (1992). Personality-intelligence relations: Assessment of typical intellectual engagement. Journal of Educational Psychology, 84, 537-552.

Heyman, G. D. & Dweck, C. S. (1992). Achievement goals and intrinsic motivation: Their relation and their role in adaptive motivation. Motivation and Emotion, 16, 231-247.

Horn, J. L., & Noll, J. (1994). A system for understanding cognitive capabilities: A theory and the evidence on which it is based. In D. K. Detterman (Ed.), Current topics in human intelligence: Vol IV. Theories of intelligence (pp. 151-204). Norwood, NJ: Ablex.

Kelley, T. L. (1927). Interpretation of educational measurement. Yonkers, NY: World Book.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working memory capacity?! Intelligence, 14, 389–433.

Marsh, H. W. (1994). Sport motivation orientations: Beware of the jingle-jangle fallacies. Journal of Sport and Exercise Psychology, 16, 365–380.

McDaniel, M. A., & Nguyen, N. T. (2001) Situational judgment tests: a review of practice and constructs assessed. International Journal of Selection and Assessment, 9, 103–113.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychological Bulletin, 114, 449-458.

Motowidlo, S. J., & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. Journal of Occupational and Organizational Psychology, 66, 337-344.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working memory and intelligence - their correlation and their relation: A commend on Ackerman, Beier, and Boyle (2005). Psychological Bulletin, 131, 61-65.

O'Connell, M. S., Hartman, N. S., McDaniel, M. A., Grubb, W. L., & Lawrence, A. (2007). Incremental validity of situational judgment tests for task and contextual job performance. International Journal of Selection and Assessment, 15, 19-29.

Odoroff, M. E. (1935). A correlational method applicable to the study of the time factor in intelligence tests. Journal of Educational Psychology, 26, 307-311.

Paterson, D. G., & Tinker, M. A. (1930). Time-limit versus work-limit methods. American Journal of Psychology, 42, 101-104.

Peak, H., & Boring, E. G. (1926). The factor of speed in intelligence. Journal of Experimental Psychology, 9, 71-94.

Rocklin, T. (1994). Relation between typical intellectual engagement and openness: Comment on Goff and Ackerman (1992). Journal of Educational Psychology, 86, 145-149.

Saucier, G. (1992). Openness versus intellect: Much ado about nothing. European Journal of Personality, 6, 381-386.

Sechrest, L. (1963). Incremental validity: A recommendation. Educational and Psychological Measurement, 23, 153-158.

Stumpf, H., Angleitner, A., Wieck, T., Jackson, D. N. & Beloch-Till, H. (1985). Deutsche Personality Research Form (PRF). Göttingen: Hogrefe.

Thorndike, E. L. (1904). An introduction to the theory of mental and social measurements. New York: Teachers College, Columbia University.

Thurstone, L. L. (1937). Ability, motivation, and speed. Psychometrika, 2, 249-254.

Trapnell, P. D. (1994). Openness versus intellect: A lexical left turn. European Journal of Personality, 8, 273-290.

Wilhelm, O., & Schulze, R. (2002). The relation of speeded and unspeeded reasoning with mental speed. Intelligence, 30, 537-554.

Wilhelm, O., Schulze, R., Schmiedek, F., & Süß, H.-M. (2003). Interindividuelle Unterschiede im typischen intellektuellen Engagement [Individual differences in typical intellectual engagement]. Diagnostica, 49, 49-60.

The author:

Oliver Wilhelm
Humboldt-Universität zu Berlin
Institute of Educational Progress
Unter den Linden 6
10099 Berlin
Germany
Tel.: +49-(0)30-2093-4873
E-Mail: oliver.wilhelm@rz.hu-berlin.de

Oliver Wilhelm is associate professor for educational assessment at Humboldt University. His research interests are focused on individual differences in cognitive abilities.

New Constructs, Methods, & Directions for Computer-Based Assessment

Patrick C. Kyllonen
Educational Testing Service, USA

Summary
The argument for computer-based testing is that it (a) promises increased efficiency and convenience, and (b) enables the assessment of new constructs using new methods not easily implemented with paper-and-pencil technology. In this paper I review a number of efficiency and convenience arguments, and focus on the merits of automatic item generation (AIG). Recent advances in AIG item development and analysis methods enable application to increasingly complex tasks, such as college-level physics, and promise continued strong future growth of AIG applications. New constructs, such as communication skills, teamwork, leadership, critical thinking, and creativity are increasingly recognized as important in both school and the workplace. Computer-based assessment methods, such as games, simulations, and reaction time methods, and tests, such as the implicit association test, and situational judgment test seem promising as ways to measure them. Situational judgment testing in particular is becoming increasingly popular in education and in industry as a way to measure a wide variety of new constructs. It is appealing to users as an authentic assessment, it can play a dual role as both an assessment and training method, and it shows less adverse impact against minority groups than other methods.

Almost from the beginning of the computer, there has been an interest in computer-based testing. Part of this interest has been based on a sense of the inevitable—anything that can be computerized will be, so why resist? But aside from inevitability there have been two major arguments for computer-based testing—convenience and the possibility of measuring constructs that cannot be measured by paper-and-pencil tests. These are the topics of this paper. The purpose of this chapter is to review a few examples of each of these topics. It is meant not to be exhaustive, but only illustrative.

Efficiency and Convenience

The case for efficiency and convenience seems straightforward. For examinees, computer-based tests can be administered 24/7 and provide quick scoring and feedback. Following testing, score reports can be transmitted rapidly to examinees and schools, employers, and other score report users. With computer-adaptive-tests, there is the additional savings of approximately half the testing time. For assessment developers, there is the convenience of rapid turnaround from item writing to administration, without mailing, shrink-wrapping, booklet production and all the other paper-and-pencil related details. For data analysts, the data entry step is eliminated, which speeds the analysis process, responding can be monitored (to avoid unnecessary mistakes), and a complete process record can be built in for auditing. For researchers and assessment designers there is the convenience of easily being able to create multiple forms and item matrix designs, and being able to randomly reorder items to mitigate sequence and position effects.

In addition, certain item types, such as speaking and writing samples, although possible with paper-and-pencil and cassette tapes, are much more easily accomplished with computers. Speaking samples collected on cassette tapes, for example, have to be marked, mailed, and sorted by hand, and tape breakage is a real possibility. Computers solve all these problems fairly easily.

Given all these advantages, why are so many major testing programs—the SAT, the Law School Admissions Tests (LSAT), the National Assessment for Educational Progress (NAEP), to name just a few – still paper-and-pencil based?

The answer is that even today there are many reasons not to do computer-based testing. Paper-and-pencil-based testing is a tried and true method; computer-based testing is still relatively new, and expensive. Computers are complicated machines that fail often and in unpredictable ways; for internet-based testing, internet traffic can cause fluctuations and disruption of service. Computers themselves are still expensive, and there is a finite supply of them. Some of the data quality advantages of computer-based testing (e.g., determining valid data, automatic fill-ins, filters and skip patterns), require considerable advance planning, and must be worked out before going to the field. Computer literacy may affect test scores, and implementation issues (e.g., font size, screen size, computer responsiveness) may too. Also, delivery software may constrain the types of questions that can be asked, perhaps requiring an awkward workaround. In general, computers add an extra layer of complication, require extra reviews, advanced setups, and tryouts, and these add expense.

There is one efficiency and convenience advantage to computer-based testing that I would like to focus on here because I think it will increase in importance in the future—automatic item generation (AIG). Strictly speaking, AIG is not limited to computer-based assessment. It can and has been used with paper-and-pencil tests (Irvine, Dann, & Anderson, 1990). But it is a technology that is greatly facilitated through the use of the computer, and it does open the possibility of being implemented in an "on-the-fly" manner, and thus can be a computer-based testing advantage.

Automatic Item Generation

Automatic item generation (AIG) is a process for having a computer write items. There are a number of different automatic item generation schemes (Irvine & Kyllonen, 2002). For example, Embretson's (1999) ART writes figural progressive matrices items by adding geometric elements together in thousands of unique ways. Systems have been developed based on grammars that allow manipulation of comparative terms (e.g., taller than, above, not) and objects (e.g., John, the dog, a square) for hundreds of different item types (e.g., Irvine, et al., 1990). ETS used such a system to create analytic reasoning items for the Graduate Record Examination (GRE; Dennis, Handley, Bradon, Evans, & Newstead, 2002). However, the most common automatic item generation scheme, and the most widely applicable, is what might be called the item-template based system. It is also the approach that is used in ETS's "Test Creation Assistant," an automatic item-generation tool (Singley & Bennett, 2002).

Slots and Fillers. The template-based system — also called a slot-and-filler system — is based on the item model (also called a form, template, schema, or shell). An item model is an item that is variabalized by a test developer. That means the test developer takes an item and turns components of the item (words, numbers) into string and integer variables (i.e., slots). The Test Creation Assistant (TCA) takes the item model, and automatically generates variants (also known as siblings, isomorphs, or clones) from it, by filling in the variabalized slots with legitimate fillers. Fillers are replacement values for the original item text and numbers. For example, consider the following mechanics item from a college-level physics examination:

> *A ball is released from rest from the top of a 200 m tall building on Earth and falls to the ground. If air resistance is negligible, which of the following is most nearly equal to the distance the ball falls during the first 4 s after it is released?*
> *(a) 40 m; (b) 80 m; (c) 120 m; (d) 200 m*

A test developer could choose to variabalize building height (e.g., 200 m vs. 400 m vs. some other height), the type of object (e.g., ball vs. rock vs. iron), or the planet (e.g., Earth vs. Mars vs. Moon) each of which would have a different gravitational formula, and so forth. Constraints would have to be specified (e.g., the key is that the distance is equal to (1/2)gt^2, where g, the acceleration due to gravity is approximately 10 m/s^2; the distance dropped would have to exceed that amount, or not if the goal was to see if the student noticed that the function was discontinuous due to the ball hitting the earth, etc.) (Kyllonen, Pfeiffenberger, Trapani, & Weng, 2009; See Bejar, Lawley, Morris et al., 2003, for an example from mathematics.)

The item model (i.e., the item marked according to which of the components serve as variables, or slots) would be entered into the Test Creation Assistant (TCA), and legitimate fillers for the various slots would also be specified. In addition, logical and mathematical constraints would be specified (e.g., that g = 10m/s^2 goes with Earth, but g = 1.6m/s^2 goes with the moon). Then, the TCA would be able to generate many item variants from a single item model, or template.

Benefits of Automatic Item Generation. There are many issues about what kinds of variables are designed to affect item difficulty, and whether they are expected to do so in a construct-relevant way or not. But one of the benefits of automatic item generation is that it forces the assessment developer to think about these very issues, which are at the core of construct validation, in a disciplined way. Another benefit of item generation is that it produces lots of items, relatively quickly and cheaply.

The Psychometrics of Automatic Item Generation. The common way to determine the quality of items is to calibrate them, that is, to determine item parameters (e.g., difficulty, discrimination) based on item-response theory. However, with automatic item generation, it is useful to calibrate item models, (i.e., families of items, variants from the same model) rather than single items, using the expected response function (ERF; the average of the item characteristic curves across item variants) (Mislevy, Wingersky, & Sheehan, 1994). In this framework an item (i.e., a variant) inherits the parameters from the item family. The benefit is that a new item, that is, a new variant, never tested, is already calibrated without pretesting based on the ERF. There is a cost to calibrating items this way. In particular, the discrimination value (i.e., slope, or item-total correlation) is lower than would be obtained from actual items, rather than variants from an item model. But the cost is relatively low. One study found that the correlation between automatically generated items and GRE scores was quite high (about r = .87; Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta, 2003).

Occasionally, some items generated automatically may be off the mark (e.g., low discrimination value; poorly predicted difficulty level), and the test developer might try to figure out why, and perhaps introduce additional constraints into the item generating process (e.g., Graf, 2008). However, a recurring finding is that the difficulty parameter is unbiased and the loss in precision is relatively small. One study (Scrams, Mislevy, & Sheehan 2002) found a loss of only 10% precision for estimating examinee ability from variants compared to actual (precalibrated) items, meaning that a computer-adaptive test based on automatic item generation would only have to be 10% longer to get the same ability estimates, with the same confidence, as a test using the original, un-cloned items.

What this means is that automatic item generation, particularly of the slot-and-filler variety, is a promising methodology that has immediate practical application. It is useful for a broad variety of tests, ranging from general ability and achievement tests to licensure tests. It has not been applied to non-cognitive assessment, but it seems reasonable to expect it to be useful for such applications as well. Continued research is underway to extend the applicability of automatic item generation schemes (e.g., Higgins, Futagi, & Deane, 2005), and to develop systems and heuristics for avoiding poorly performing items, or statistically adjusting for poorly performing items once they are created, but thus far the evidence suggests that the methodology is useful, and even weird items are not that consequential in obtaining accurate scores.

New Item Types

Almost from the beginning of the promotion of computer-based testing to the using community, there were concerns that the benefits of increased efficiency and decreased examinee testing time were not sufficient to outweigh the costs and risks of converting tests to computer-based assessments (e.g., Sands, Waters, & McBride, 1997). The argument was made that what made computer-based testing worthwhile, what put the benefit/cost ratio into a favorable status, was the additional opportunity to measure abilities that simply could not be measured with paper-and-pencil tests. In the remainder of the paper we consider what those opportunities might be.

New Constructs

A recent survey of 400 U.S. employers identified several applied skills as “very important” for success in the 21st century workforce (Conference Board et al., 2006). A finding was that these applied skills, or new constructs, were rated as important as or more important than traditional content skills, such as writing, reading, and math. These applied skills included both cognitive factors: critical thinking, creativity, communications skills, and technological proficiency; and non-cognitive factors: work ethic, leadership, teamwork, and ability to work in a diverse workplace setting. At the same time employers mentioned that most graduates of secondary and postsecondary educational institutions were not well prepared with respect to these applied skills. An issue in assessing these kinds of new constructs is that traditional methods, such as paper-and-pencil self-reports, are not necessarily the most valid approaches for measuring them.

New Methods

Computer-based assessment opens new possibilities for measuring a range of new constructs, including the so-called 21st century skills (Conference Board et al., 2006), and also ones such as those measured by the new science assessments using interactive computer testing methods in both PISA and the National Assessment of Educational Progress (NAEP). Simulations and games are promising new computer-based assessment methods. But in this paper I limit the focus to three methods — reaction time, implicit association tests, and situational judgment tests. These are new methods in the sense that computer-based assessment and scoring makes these approaches feasible in a way that previous methods (e.g., personalized testing) were not.

Reaction Time. Time to respond to presented stimuli, such as test items, is difficult to measure with paper-and-pencil, albeit awkwardly possible with speeded tests, but easy to measure with computers. Response speed itself is a basic cognitive ability, and separate from speed of responding on paper-and-pencil tests (Carroll, 1993). However, the importance of cognitive speededness per se has not been demonstrated for educational achievement or other broadly important real-world outcomes. Partly this may be due to measurement problems, most notably that response time on an item reflects both an ability — cognitive speed — and a choice or style — how long to persist before quitting the item and moving to the next one. There is a resurgence of interest in these issues and the topic of response speed (e.g., Van der Linden, 2007), and so there may be significant developments soon regarding cognitive speed in the context of achievement testing.

Implicit Association Test. However, measuring cognitive speededness per se is not the only application involving measuring reaction time. The implicit association test (IAT) is a method of using reaction time to measure non-cognitive factors that seems particularly promising. The IAT is already widely used in social psychology research to measure attitudes and preferences (e.g., Greenwald, Poelman, Uhlmann, Banaji, in press). Although still at the research stage, the IAT seems to be a powerful method for measuring non-cognitive factors that are not easily measured with simple self-assessments or ratings by others. It works by comparing an examinee’s response time to different kinds of stimulus pairs, with the faster reaction indicating which pair is more natural in the mind of the examinee. Thus, the method can uncover stereotypical attitudes (e.g., gender or racial bias) or preferences (e.g., political candidates), and applications to measure personality are starting to emerge (Schnabel, Banse, & Asendorpf, 2006).

Situational Judgment Tests. These involve presenting a scenario involving a problem and asking the examinee for the best way to solve the problem. The presentation can be written, audio, or video (animation or live actors), and the response can be likewise, or it can involve an open-ended written or spoken response, along with a justification. Consider the following example we recently designed to measure teamwork, specifically, resolving conflicts and negotiating (Zhuang, MacCann, Wang, Liu, & Roberts, 2008):

> *You have recently formed a study group with several of your classmates in order to prepare for a difficult final exam. Unfortunately, the various members of the group have very different schedules, so you all meet after class one day to try to work out a final schedule for your group review sessions.*
>
> *Which of the following is the most important factor to consider in weighing any proposed suggestions?*
>
> *(a) Making sure that the schedule will allow the smartest students to attend, so that the study group will cover more material.*
>
> *(b) Making sure the proposed meeting times do not conflict with your own course schedule.*
>
> *(c) Yielding to the majority of the group even if it means some members will not be able to participate.*
>
> *(d) Breaking the group down into sub-groups based on compatible schedules **

A meta-analysis (McDaniel, Hartman, Whetzel, & Grubb, 2007) showed that these kinds of situational judgment items measure both cognitive ability (r = .43) and personality, particularly conscientiousness (r = .33) and agreeableness (r = .20), and that they have predictive validity with respect to real-world outcomes (r = .33). If instead of asking for the best response, the respondent is asked what he or she would do, situational judgment tests become more like personality measures (r = .51, .53 for conscientiousness and agreeableness) and less like ability measures (.23), while still predicting outcomes (r = .27).

Situational judgment tests (SJTs) are a method, and likely with some diligence and creativity, a method useful for assessing a wide variety of constructs ranging from communication skills to work ethic, teamwork, leadership, and other factors identified in the Conference Board et al. (2006) survey. The basic methodology is to interview subject matter experts within a domain (e.g., pilots, medical doctors, teachers, jet engine mechanics, school principals), identify critical incidents that illustrate the particular construct targeted (e.g., challenging communications situations; situations in which leadership was called for), and then to write items based on those critical incidents. The advantage of SJTs over traditional content items is that SJTs can reflect judgment calls, when the situation is not clear cut and there is no obvious correct answer. It is these judgment calls that often represent true expertise in a domain. Thus keying the correct answer is a challenging task in itself, and there is a literature on various ways to do that (e.g., Legree & Psotka, 2006). Also, as a consequence of SJTs being essentially judgments, reliability is typically lower than for fact-based items, such as vocabulary or arithmetic items. Testing time consequently can be two or three times longer than content-based items to achieve comparable reliability.

But SJTs present many advantages over traditional measures. Because the situations are drawn from the practical experiences of students or job incumbents, SJTs have high content validity, and can double as training or job previews and as training materials. They potentially can reflect qualities that may be difficult to measure with traditional paper-and-pencil tests, and they are well suited to passing the test of authenticity. Video SJTs seem particularly promising in this regard (e.g., Olson-Buchanon & Drasgow, 2006). In addition, research has demonstrated that they show less adverse impact against minority groups, and some applications in education suggest that they add to cognitive ability measures in predicting important outcomes such as finishing school. Due to all these factors, SJTs are in high demand and becoming an increasingly popular means for assessing non-cognitive factors (Kyllonen & Lee, 2005).

Summary

From the beginning of computer-based testing, the argument for it has been that it (a) promises increased efficiency and convenience, and (b) enables the assessment of new constructs using new methods not easily implemented with paper-and-pencil technology. In this paper I reviewed a number of efficiency and convenience arguments, and expanded on the merits of automatic item generation (AIG). AIG is a powerful technology for increasing the production of new items, which is important to satisfy the item demand created by the increase in testing in general and to satisfy the item needs of computer-adaptive testing in particular. AIG also provides a disciplined means to focus attention on the central issue of construct validity. Advances in both item development and analysis methods associated with AIG enable application to increasingly complex tasks, such as college-level physics, and suggest continued strong future growth of AIG applications.

Assessment of new constructs is a second area in which the benefits of computer-based testing will be realized. New constructs, such as communication skills, teamwork, leadership, critical thinking, and creativity are increasingly recognized as important in both school and the workplace. Computer-based assessment methods seem promising as ways to measure new constructs. The assessment of cognitive processing time has been a long-standing promise of computer-based testing, but progress has been slow. New statistical and psychometric methods may finally provide breakthroughs in this area, and additionally hold promise in measuring personality. One method for measuring attitudes and preferences, the implicit association test seems particularly promising. Situational judgment testing is becoming increasingly popular in education and in industry as a way to measure a wide variety of new constructs. It is appealing to users in that it represents an authentic assessment, it can play a dual role as both an assessment and training method, and it shows less adverse impact against minority groups than other methods.

References

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

Carroll, J. B. (1993). Human cognitive abilities. New York: Cambridge University Press.

Conference Board, Partnership for 21st Century Skills, Corporate Voices for Working Families, & Society for Human Resources Management (2006). Are they really ready to work? Employers perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce. New York: The Conference Board.

Dennis, I., Handley, S., Bradon, P., Evans, J. & Newstead, S. (2002). Approaches to Modeling Item-Generative Tests. In S. H. Irvine & P.C. Kyllonen, (Eds.), Item generation for test development. Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, S.E. (1999). Generating items during testing: Psychometric issues and models. Psychometrika, 64, 407-433.

Graf, A. (2008). Approaches to the Design of Diagnostic Item Models. Technical Report Number ETS RR-08-07. Princeton, NJ: Educational Testing Service.

Greenwald, A.G., Poelman, T.A., Uhlmann, E.L., & Banaji, M.R. (in press). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. Journal of Personality and Social Psychology.(retrieved November 25, 2008, from http://faculty.washington.edu/agg/pdf/UUIAT3.Complete.30Oct08.pdf).

Higgins, D., Futagi, Y. & Deane, P. (2005). Multilingual Generalization of the Model Creator Software for Math Item Generation ETS Research and Development (ETS RR-05-02) Princeton, NJ: ETS.

Irvine, S. H. & Kyllonen, P.C. (2002). Item generation for test development. Mahwah, NJ: Lawrence Erlbaum Associates.

Irvine, S.H., Dann, P.L., & Anderson, J.D. (1990). Towards a theory of algorithm-determined cognitive test construction. British Journal of Psychology, 81, 173-195.

Kyllonen, P. C., Pfeiffenberger. W, Trapani, C., & Weng, P. (2009). Evaluating Transfer Learning in College-Level Physics: Final Report. Princeton, NJ: Educational Testing Service.

Kyllonen, P. C., & Lee, S. (2005). Assessing problem solving in context. In O. Wilhelm & R. W. Engle (Eds.) *Handbook of Understanding and Measuring Intelligence* (pp. 11-25). Thousand Oaks, CA: Sage.

Legree, P. & Psotka, J. (2006). *Refining situational judgment test methods.* In Proceedings of the 25th Army Science Conference. Orlando, FL.

McDaniel, M.A., Hartman, N.S., Whetzel, D.L. & Grubb. W.L., III (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology, 60,* 63-91.

Mislevy, R.J., Wingersky, M.S., Sheehan, K.M. (1994). *Dealing with uncertainty about item parameters: Expected response functions.* ETS Research Report Number 94-28-ONR. Princeton, NJ: Educational Testing Service.

Olson-Buchanon, J.B. & Drasgow, F. (2006). Multimedia situational judgment tests: The medium creates the message. In J.A. Weekley & R.E. Ployhart (Eds.), Situational Judgment Tests (pp. 253-278). Routledge.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Schnabel, K., Banse, R., & Asendorpf, J. (2006). Assessment of implicit personality self-concept using the implicit association test (IAT): Concurrent assessment of anxiousness and angriness. *British Journal of Social Psychology, 45 (2),* 373-396.

Scrams, D.J., Mislevy, R.J., & Sheehan, K.M. (2002). *An Analysis of Similarities in Item Functioning Within Antonym and Analogy Variant Families.* Technical Report Number ETS RR-02-13. Princeton, NJ: Educational Testing Service.

Singley, M.K., & Bennett, R.E. (2002). Item Generation and Beyond: Applications of Schema Theory to Mathematics Assessment. In S. H. Irvine & P.C. Kyllonen, (Eds.), *Item generation for test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72(3),* 287-308.

Zhuang, X., MacCann, C., Wang, L, Liu, L., & Roberts, R.D. (2008). *Development and Validity Evidence Supporting a Teamwork and Collaboration Assessment for High School Students.* Technical Report Number ETS RR-08-50. Princeton, NJ: Educational Testing Service.

The author:
Patrick C. Kyllonen
Educational Testing Service
Rosedale Road
Princeton, NJ 08541
USA

E-Mail: pkyllonen@ets.org

Patrick Kyllonen is the Director of the New Constructs centre at ETS. The centre is responsible for background variable development for the U.S.'s National Assessment of Educational Progress (NAEP), which assesses the achievement of over a million students every year on a wide variety of subjects ranging from reading, math, and writing, to science, social studies, art, geography, and computer literacy. In addition to NAEP, the centre conducts research on non-cognitive assessment for K-12, higher education, and workforce applications, and is introducing a new high-stakes non-cognitive assessment for graduate school admissions beginning July 2009 as part of the Graduate Record Examination (GRE).

Measuring Complex Problem Solving: The MicroDYN approach

Samuel Greiff & Joachim Funke
University of Heidelberg, Germany

Abstract
In educational large-scale assessments such as PISA only recently an increasing interest in measuring cross-curricular competencies can be observed. These are now discovered as valuable aspects of school achievement. Complex problem solving (CPS) describes an interesting construct for the diagnostics of domain-general competencies. Here, we present MicroDYN, a new approach for computer-based assessment of CPS. We introduce the new concept, describe proper software and present first results. At last, we provide an outlook for further research and specify necessary steps to take in the effort to measure CPS on an individual level.

Until recently, psychological assessment of aptitudes and abilities has relied almost entirely on paper-and-pencil-based testing. As computers emerged, these were discovered as efficient means to measure abilities. This development has led to new technologies and assessment procedures such as Computer Adaptive Testing (CAT) as is outlined widely in this volume. However, not only has measurement become more efficient through computer-based assessment. Additionally, new constructs not measurable in traditional formats now can be assessed by computer-based procedures (see Patrick Kyllonen, this volume). Among others, complex problem solving being inherently dynamic is one of these new constructs that rely on interaction between task and subject. We will introduce complex problem solving as research topic and present ways to measure problem-solving competencies in an innovative way. First results and open-access software are presented showing how new constructs over time might emerge.

Complex problem solving within dynamic systems has been an area of major interest in experimental research over the last decades (for a review see Blech & Funke, 2005). Comparatively little research has been conducted about CPS in the context of individual differences even though some efforts have been made (e.g. Beckmann, 1994; Wagener, 2001). However, embedded in the recent development of large-scale assessments in educational settings, cross-curricular competencies such as CPS have been discovered as valuable aspects of school achievement (Klieme, Leutner, & Wirth, 2005).

Starting from a practical point of view, applied implications of CPS are frequently found in everyday life. Many activities can be described within this formal framework ranging from medical emergencies over evaluating one’s monthly expenses to handling ticket machines at train stations. These activities involve situations comprising of the following characteristics:

- Different variables influence one or more outcomes (interconnectedness),
- the underlying system is not static (dynamics),
- exhaustive information and evaluation of the situation may not be obtained (intransparency).

A first successful approach towards measuring CPS (CPS and dynamic problem solving are identical; we argue that CPS is in itself always dynamic as opposed to analytical problem solving) in a large-scale context was conducted in PISA 1999 (Wirth & Funke, 2005). Students had to explore and control a system (embedded in the context of space travel) with different states that could be changed by activating or deactivating various switches (e.g. on/off; start/land). A system with qualitatively differing states that can be altered by the user is commonly called a finite automaton. Comparable to a finite automaton is the approach outlined below. These systems differ, however, from the qualitative approach by using only quantitatively different states (e.g. continuum from low to high). The finite automaton used in PISA could explain additional variance in student achievement after controlling for general intelligence. Furthermore, factor analytical results, structural equation models and multidimensional scaling suggested that CPS, analytical problem solving, domain specific literacy and general intelligence are correlated and yet separable constructs with CPS being best separable from the others (Wirth, Leutner, & Klieme, 2005).
These results indicate construct validity and in particular convergent and divergent validity for CPS. However, the finite automaton used in

PISA was an ad hoc constructed instrument with questionable psychometric qualities so that measurement range and classification remains unclear calling for a properly piloted and validated testing device. A new approach is outlined in this paper and first empirical results are presented. Milestones on the way to measuring CPS are further specified.

The MicroDYN Approach

Despite the awakening interest in individual differences, there is still a substantial lack of well-scrutinized testing devices. Additionally, little agreement on how to measure CPS on an individual level has been reached and sound theoretical foundations to be used as starting points are still rare (Greiff & Funke, 2008b).

Another major shortcoming of complex problem-solving research as it was introduced by Dörner in the 1970s (Funke & Frensch, 2007) is its "one-item-testing". Virtually all devices consist of one large and rather complicated scenario the participant has to work through. At the end either overall performance or various status and process indicators are calculated and evaluated. Thus, CPS instruments are tests, which contain exactly one excessive item, or at best one item bundle speaking in IRT-terms (Embretson & Reise, 2000) if various independent subsystems are considered as some authors do (e.g. Müller, 1993). Other tests allow subjects to explore a given system over a period of time and then ask several questions about this one system. That does not make the answers any less dependent.

Bearing these severe limitations in mind, the question arises how dynamic problem solving could possibly be measured with psychological tests. We assume that individual differences might possibly be detected within the formal framework of linear structural equation systems (LSE-systems), which we call the MicroDYN approach. This type of items has been used considerably in experimental research as indicator for problem solving performance (Blech & Funke, 2005). The basic approach here, however, is now a different one.

Items based on this approach require participants to detect causal relations and control the presented systems. We suppose that the everyday examples mentioned above can be modelled by MicroDYN systems since advanced skills in strategic planning, internal model building and system control are crucial in the specified situations as well as tested within the framework of MicroDYN systems. To solve the severe problem of one-item-testing, various completely independent systems are presented to the subjects (see below).

To summarize, we choose to work within the formal framework of linear structural equation systems. The MicroDYN approach may be able to overcome some of the shortcomings mentioned above:

1. The lack of sound theoretical frameworks calls for a different kind of framework, which MicroDYN systems offer formally (theoretical embedment).
2. MicroDYN systems are easily constructed and can be varied in difficulty freely (infinite item pool).
3. A sufficient number of divergent items can be presented (item independency).
4. Many everyday activities can be described by MicroDYN items (ecological validity).

The Items

An example of a typical MicroDYN item is presented in Figure 1. MicroDYN systems consist of exogenous variables, which influence endogenous variables, where only the former can be actively manipulated. Possible effects include main effects, multiple effects, multiple dependencies, autoregressive processes of first order ("eigendynamics"), and side effects, which all can be freely combined.

Main effects describe causal relations from exactly one exogenous variable to exactly one endogenous variable. If an exogenous variable is involved in more than one main effect, this is labelled a multiple effect. Effects on an endogenous variable influenced by more than one exogenous variable are labelled multiple dependence. Participants can actively control these three effects as they manipulate the values of exogenous variables within a given range. Effects merely incorporated within endogenous variables are called side effects when endogenous variables influence each other, and eigendynamics when endogenous variables influence themselves (i.e. growth and shrinkage curves) due to the dynamic a variable develops by itself as time passes (e.g. bacteria cultures). Participants cannot influence these two effects directly; however, they are detectable by adequate use of strategy.

Additionally, all effects may differ in path strength.

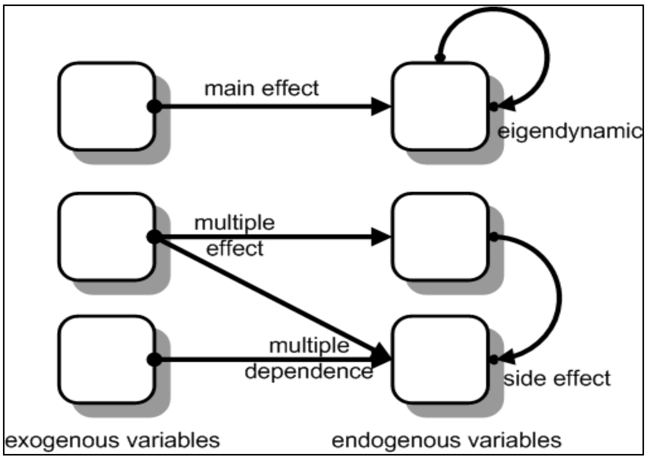


Figure 1: Underlying structure of a MicroDYN item with all possible effects displayed

Participants face between 10 and 12 of these items each lasting about 5 minutes summing to an overall testing time of approximately one hour including instruction and trial time. The MicroDYN items are minimally but sufficiently complex and at the same time adequately in number. Each item is processed in three stages:

- Stage 1, exploration phase: Participants can freely explore the system. No restrictions or goals are presented at this time apart from getting acquainted with the system and the way it works. Participants can reset the system or undo their last steps. A history to trace prior steps is provided. Exploration strategies can thus be assessed.

- Stage 2, drawing the mental model: Simultaneously (or subsequently) to their exploration, participants are asked to draw the connections between variables as they suppose. This helps in assessing acquired causal knowledge (declarative knowledge is tested)

- Stage 3, control phase: Participants are asked to reach given target values on the endogenous variables by entering adequate values for the exogenous variables. During this phase, the practical application of the acquired knowledge is assessed (procedural knowledge is tested).

Current Research

Up to now little knowledge exits about how MicroDYN systems behave and which attributes cause their difficulty despite their extensive use in experimental research in the last decades. Based on a detailed task-analysis, seven factors are identified as potentially relevant for item difficulty (Table 1).

Testing these item-characteristics is understood as a first step to competence levels. The research design, first result and a brief discussion are provided below.

(1)	**Quality of effects**	Different causal relationships (as depicted in figure above)
(2)	**Quantity of effects**	Number of effects (regardless their quality)
(3)	Strength of paths	Specifies strength of an effect (and hence its detectability)
(4)	**Number of variables**	Mere number of exogenous and endogenous variables
(5)	Variable dispersion	Specifies how closely a given number of effects clusters on the variables
(6)	Effect configuration	Order and alignment
(7)	Starting & target values	Self-explaining; target values influence only endogenous variables

Table 1: Attributes potentially determining difficulty in MicroDYN systems and their explanation.

Design

We used a within-subject design (n=50) with repeated measures on all factors. An overall of 15 MicroDYN systems was presented, each lasting about 5 minutes (split on two sessions). The independent variables mainly focused were Quality of effects, Quantity of effects and Number of variables (bold in Table 1).

Quality of effects: Main effects, multiple effects and side effects were tested against each other as can be seen in Figure 1 (multiple dependencies and eigendynamics were not tested at this stage).

Quantity of effects: Two different quantities (2 vs. 4 effects) were tested against each other. This is outlined schematically in Figure 2.
Number of variables: Systems were constructed equally only differing in number of variables as can be seen from Figure 3.

Dependent variables

Correctness of mental model: Subjects are asked to draw the connections between variables as they suppose. Better performance is indicated by a higher value on the dependent variable. The difference between correctly and incorrectly drawn connections in relation to the total number of correct connections was used to indicate performance.

Control performance: After exploring the system extensively, subjects are asked to reach given target values on the endogenous variables as control task (results not yet available).

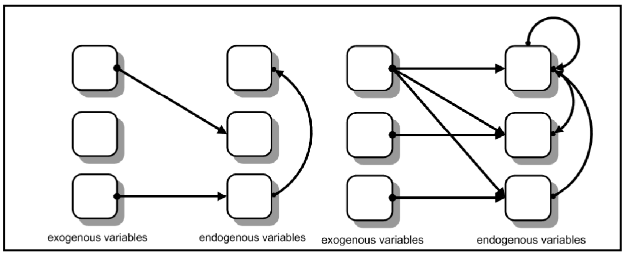


Figure 2: Two items with low resp. high number of effects.

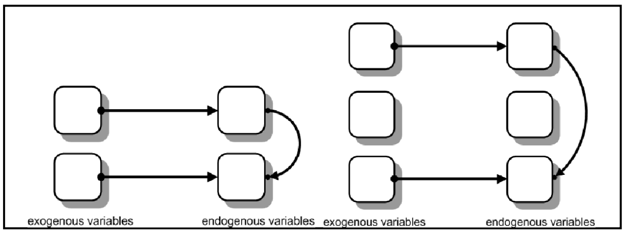


Figure 3: Two items with 2 resp. 3 exogenous and endogenous variables.

Results

Table 2 provides an overview of the ANOVA-results. There is a medium strong effect for Number of variables indicating that two systems being totally equal the one with more additional (and unnecessary) variables is more difficult. The explained variance is 0,16. A graphical depiction is found in Figure 4.

Independent variable	F	df_{Num}	df_{Denom}	p	Eta^2 (partial)
Number of exogenous & endogenous Variables	*8,650*	*2*	*92*	*0,001***	*0,158*
Quality of effects	*18,270*	*2*	*90*	*0,001***	*0,289*
Quantity of effects	2,290	1	45	>0,10	0,048
Quality x Quantity	0,500	2	90	>0,05	0,011

Table 2: ANOVA results for the tested effects.

There is a strong effect for Quality of effects showing that side effects increase difficulty heavily. This might be because side effects can only be observed but not actively manipulated. Multiple effects and main effects do not vary significantly in the dependent variable (contrast not shown); however, multiple effects seem to be slightly easier. This might be due to participants' a priori expectation of a higher likelihood for multiple effects as these occur most frequently in real world settings. The explained variance is 0,29. A graphical depiction is found in Figure 5.

Surprisingly, items with only 2 effects are not easier than those having 4 effects. Apparently, the opposite might be true even though not statistically reliable. This unexpected result might be due to problems with the dependent variable we chose as outlined below. The explained variance is 0,05 and non-significant. A graphical depiction is found in Figure 5.

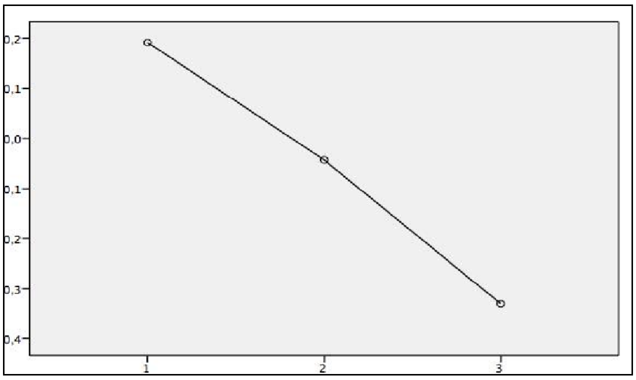


Figure 4: Effects of *Number of variables* on the correctness of the mental model. Ordinate: performance. Abscissa: Number of exogenous and endogenous variables (ranging from 2 to 4).

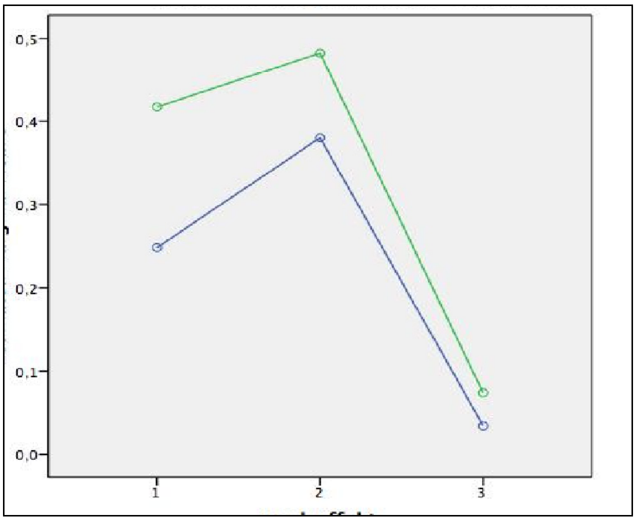


Figure 5: Effects of *Quality and Quantity of effects* on the correctness of the mental model. Ordinate: performance; Abscissa: Quality of effects (1=main effect, 2=multiple effect, 3=side effect); light line: 4 effects, dark line: 2 effects.

There is no interaction between Quality and Quantity of effects. Other interactions were not planned in the design.

Further screening of the data suggests the following effects:

- There is some evidence for problems with the dependent variable. These might be overcome by more complex indicators. Currently, a simulation study is carried out to decide which indicators represent problem-solving performance best.

- Correctness of mental model and control performance are weakly correlated (averaged r=0.15) suggesting that results might look differently for control performance.

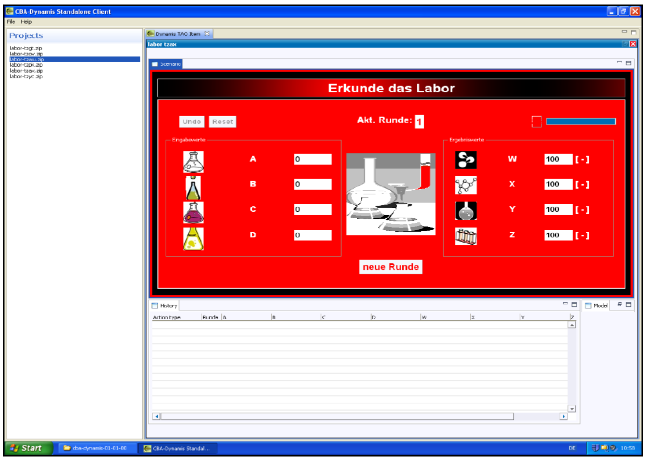
- Subjects have considerable problems detecting side effects and tend to mistake them as two- to four-way multiple effects.

- There are only moderate training effects. As time passes, subjects perform slightly better. However, the training effect is less than half a standard deviation.

Implementation

The programming and development of the software is carried out in close cooperation with the DIPF (Frankfurt, Germany) and SOFTCON (Munich, Germany). The final version will leave considerable freedom to the researcher regarding graphical layout, semantics and item generation.

Currently, the software is in the process of development. It runs stable in a preliminary version. An authoring tool integrated in the open-access platform TAO (Plichart, Jadoul, Vandenabeele, & Latour, 2004; Reeff & Martin, in press) will be released late 2008/early 2009. An up-to-date screenshot is presented in Figure 6.



A 0
B 0
C 0
D 0
W 100 [-]
X 100 [-]
Y 100 [-]
Z 100 [-]

Figure 6: Screenshot of the MicroDYN software.

In the left panel loaded and ready-to-start items are displayed. The red box is the actual item consisting of exogenous variables on the left and endogenous on the right. Additionally, an elapsed-time meter, a round counter, a reset and an undo-button are available. The history is placed at the page bottom. Here participants can trace their former manipulations and their effects for deeper analysis.

Perspective

Data acquisition for the first experiment finished in August 2008. Data have been presented recently on two conferences (Greiff & Funke, 2008a, 2008b); in-depth analyses are currently carried out.

There is need for a follow-up study to learn more about item difficulty (i.e. multiple dependencies and eigendynamics have yet not been studied) in MicroDYN systems, which will start within the next weeks. Subsequently, explorative competence levels can be derived and tested in a pilot study. Simultaneously, the existing software is upgraded. The preliminary time schedule is shown in Figure 7.

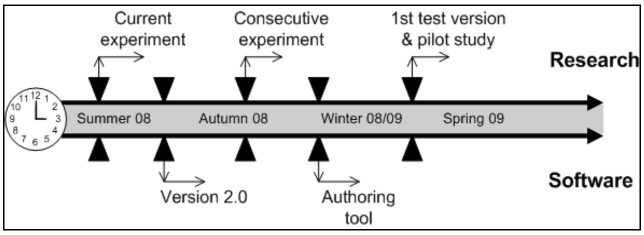


Figure 7: MicroDYN development: Preliminary time schedule until middle 2009.

Not yet incorporated are aspects of strategy and process data. By looking at the way subjects explore a system, different strategies can be identified and evaluated. This promising approach has been widely neglected in psychological diagnostics so far and is a promising field of enhancing prediction in achievement facets. First interesting ideas can be found in Rollett (2007).

The aim of the MicroDYN approach is to provide a well-scrutinized and empirically valid testing instrument for dynamic problem solving, which covers cognitive facets that yet cannot be tested by conventional tests of cognitive ability.

Applicability and Perspective

If CPS can be nomothetically classified and established as a valid construct it might be relevant in virtually all areas involving prediction or explanation of cognitive performance.

In the context of educational large-scale assessments, a detailed analysis of factors determining difficulty as described yields important information for item construction and is a prerequisite for a formally and theoretically valid testing device for individual competence levels in CPS.

MicroDYN might capture a construct yet not testable in cognitive psychology. Testing subjects on independent items in dynamic and interactive situations looking simultaneously at process and status data opens new doors in prediction of performance in various cognitive constructs such as student achievement.

However, various obstacles related to the computerized testing environment as well as theoretical questions must be overcome. Technically, a test that is administered via the internet must run stable with different local networks and on varying hardware. Experience shows that technical issues of computer-based testing are usually (too) easily disregarded.

From a theoretical point of view, a construct - however measured - must be theoretically grounded and should yield indicators for various performance aspects. Existing problem solving theories are unspecific and not sufficiently validated as to allow their use in test development. Thorough technical planning and theoretical research is needed to deal with these obstacles adequately.

In summary, CPS is seen as a key qualification for success in life. For this reason, it receives interest from large-scale assessment studies like PISA or PIAAC. The growing interest in problem solving increases the need for efficient assessment procedures. One promising approach is outlined in this paper.

Acknowledgements

The preparation of this paper was supported by grant Fu 173/11-1 from the German Research Foundation (DFG) in the Priority Programme “Models of Competencies for Assessment of Individual Learning Outcomes and the Evaluation of Educational Processes” (SPP 1293).

References

Beckmann, J. F. (1994). Lernen und komplexes Problemlösen. Ein Beitrag zur Konstruktvalidierung von Lerntests [Learning and complex problem-solving. A contribution to validate the construct of learning tests]. Bonn: Holos.

Blech, C., & Funke, J. (2005). Dynamis review: An overview about applications of the Dynamis approach in cognitive psychology. Bonn: Deutsches Institut für Erwachsenenbildung. Electronically available under http://www.die-bonn.de/esprid/dokumente/doc-2005/blech05_01.pdf

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum.

Frensch, P. A., & Funke, J. (Eds.). (1995). Complex problem solving: The European perspective. Hillsdale, NJ: Lawrence Erlbaum.

Funke, J., & Frensch, P. A. (2007). Complex problem solving: The European perspective - 10 years after. In D. H. Jonassen (Ed.), Learning to solve complex scientific problems (pp. 25-47). New York: Lawrence Erlbaum.

Greiff, S., & Funke, J. (2008a). Schwierigkeiten in Problemlöseszenarien - Was ist das und was macht sie aus? [Difficulty in problem-solving scenarios – what it is and how it is determined] Paper presented at the AEPF, Kiel, 26th August 2008.

Greiff, S., & Funke, J. (2008b). What makes a problem complex? Factors determining difficulty in dynamic situations and implications for diagnosing complex problem solving competence. In J. Zumbach, N. Schwartz, T.

Seufert & L. Kester (Eds.), Beyond knowledge: the legacy of competence (pp. 199-200). Wien: Springer.

Klieme, E., Leutner, D., & Wirth, J. (Eds.). (2005). Problemlösekompetenz von Schülerinnen und Schülern. [Problem solving competency of students] Wiesbaden: VS Verlag für Sozialwissenschaften.

Müller, H. (1993). Komplexes Problemlösen: Reliabilität und Wissen [Complex problem solving: Reliability and knowledge]. Bonn: Holos.

Plichart, P., Jadoul, R., Vandenabeele, L., & Latour, T. (2004). TAO, a collaborative distributed computer-based assessment framework built on Semantic Web standards. Paper presented at the AISTA, Luxembourg.

Reeff, J.-P., & Martin, R. (in press). Use of the internet for the assessment of students' achievement. In J. Hartig, E. Klieme & D. Leutner (Eds.), Assessment of competencies in educational settings. Göttingen: Hogrefe & Huber.

Rollett, W. (2007). Strategieeinsatz, erzeugte Information und Informationsnutzung bei der Exploration und Steuerung komplexer dynamischer Systeme. Dissertationsschrift. [Use of strategy, generated information and use of information when exploring and controlling complex dynamic systems] Braunschweig: Technische Universität Carolo-Wilhelmina.

Wagener, D. (2001). Psychologische Diagnostik mit komplexen Szenarios. Taxonomie, Entwicklung, Evaluation [Psychological Diagnostics with complex scenarios. Taxonomy, development, evaluation]. Lengerich: Pabst Science Publishers.

Wirth, J., & Funke, J. (2005). Dynamisches Problemlösen: Entwicklung und Evaluation eines neuen Messverfahrens zum Steuern komplexer Systeme [Dynamic Problem solving: Development and evaluation of a new measuring device to control complex systems]. In E. Klieme, D. Leutner & J. Wirth (Eds.), Problemlösekompetenz von Schülerinnen und Schülern (pp. 55-72). Wiesbaden: VS Verlag für Sozialwissenschaften.

Wirth, J., Leutner, D., & Klieme, E. (2005). Problemlösekompetenz - Ökonomisch und zugleich differenziert erfassbar? [Problem-solving competency – can it be measured economical and differentiated?] In E. Klieme,

D. Leutner & J. Wirth (Eds.), Problemlösekompetenz von Schülerinnen und Schülern [Problem-solving competence in students] (pp. 73-82). Wiesbaden: VS Verlag für Sozialwissenschaften.

The authors:

Samuel Greiff
Department of Psychology
University of Heidelberg
D-69117 Heidelberg, Germany
+49 6221 54 7613
E-Mail: samuel.greiff@psychologie.uni-heidelberg.de

Joachim Funke
Department of Psychology
University of Heidelberg
D-69117 Heidelberg, Germany
+49 6221 54 7305

E-Mail: Joachim.Funke@urz.uni-heidelberg.de
WWW: http://www.psychologie.uni-heidelberg.de/ae/allg/forschun/dfg_komp/index.html

Samuel Greiff is a research assistant at the Department of Psychology, University of Heidelberg. He studied in Marburg (Germany), Bergen (Norway) and Heidelberg, where he received his master of psychology in 2006. He teaches several subjects within the area of quantitative psychometrics and general psychology for undergraduate and graduate students at Heidelberg University and at the College of Heidelberg. His main research interests include test development and interindividual differences in complex problem solving. Currently, he is developing a psychometrically sound test for measuring dynamic problem solving ability that is administered fully computer-based.

Joachim Funke, Ph.D., is a full Professor of Psychology at the University of Heidelberg, where he chairs the Unit of General and Theoretical Psychology. He has conducted empirical research on complex problem solving for many years and is considered the founder of research on dynamic systems in Germany. In 2003 he developed a finite automaton ("space travel") that measured problem solving ability in PISA 2003. Additional research interests are within emotion and cognition, complex cognition and creativity. Since last year, Dr. Funke is a member of the Marsilius fellowship.

Testing for Equivalence of Test Data across Media

Ulrich Schroeders
Humboldt-Universität zu Berlin, Germany

Abstract:
In order to stay abreast of social and technological changes and to capitalize on potential advantages of computer-based over paper-pencil testing, researchers are – in a first step of this transition process – concerned with moving already existing psychometric measures to computers. Therefore, testing for equivalence of a measure across test media becomes important in understanding whether computerizing measures affect the assessment of the underlying construct positively or adversely. In practical terms during the transition period equivalence is crucial in order to guarantee the comparability of test data and therewith the fairness for the test takers across media. After defining the term equivalence the available empirical evidence and proper statistical methods are discussed. It is argued that confirmatory factorial analysis is the soundest statistical tool for equivalence comparisons. The chapter concludes with some practical advices on what to do in order to adequately support the claim that a measure is equivalent across test media.

Given the potential advantages of computer-based testing (CBT) over paper-pencil-testing (PPT) – like computer adaptive testing (CAT, see Thompson & Weiss, this volume) or the potential to reduce costs of testing (see Latour, this volume) – educational and psychological testing is transferred more frequently to a new test medium. Besides large scale studies (e.g., NAEP, see Bridgeman or CBAS, see Haldane, both this volume) there is a variety of small scale studies. In an initial step researchers are concerned to transfer already available paper-based measures to computers. Subsequently, opportunities provided by the new test medium like multimedia extensions might catch a researcher’s interest and trigger changes of the instrument with regard to its content. These two trends reflect two different research strategies. This chapter addresses data analytic issues within the first research strategy, primarily, the issue of equivalence of measures across test media. It is divided into three sections: (A) What is equivalence?, (B) Is there equivalence?, and (C) How to test for equivalence? The chapter concludes with some practical recommendations on how to achieve equivalence.

What is equivalence?

Searching for the term equivalence in the "Standards for educational and psychological testing" (AERA, APA, & NCME, 1999) you will find several passages dealing with the issue. In the paragraphs about test administration (p. 62), score comparability (p. 57), and fairness of testing (p. 73) equivalence is immanent, but could easily be replaced by different labels like "unbiasedness" or "test fairness". We will use the term "equivalence" following this working definition: *The scores of measures are equivalent if they capture the same construct with the same measurement precision, providing interchangeable scores for individual persons.* This definition suggests that two measures are equivalent if they are strict parallel, that is, test scores of such measures are solely dependent on the latent ability dimension and independent of test administration. Equivalence is given if the underlying source of all within group variance also accounts for the complete variance between the groups (PP vs. PC). Thus, equivalence is accurately described as measurement invariance (Drasgow & Kanfer, 1985). As we will see later on, there are different types of equivalence or measurement invariance. The next section will shed some light on the question whether evidence for equivalence can be found in the literature of educational and psychological testing.

Is there equivalence?

Numerous studies try to clarify the question of equivalence across test media with respect to a) a specific measure (e.g. the computerized GRE, Goldberg & Pedulla, 2002), b) specific subgroups of testees (e.g. ethnic or gender groups, Gallagher, Bridgeman, & Cahalan, 2002) or c) specific soft- and hardware realizations (e.g. pen-based computer input, Overton, Taylor, Zickar, & Harms, 1996). However, the findings of these studies often remained unconnected and inconclusive. Mead and Drasgow (1993) attempted to connect these individual findings in their frequently cited – but by now outdated – meta-analytical study. Their synopsis endorses the structural equivalence of ability test data for power tests gathered through CBT versus PPT. The cross-mode correlation corrected for measurement error was r = .97 whereas this coefficient was only r = .72 for speed tests. The authors argue that the reason for the low cross-mode correlation among speed tests is substantiated in different motor skill requirements and differences in presentation (instructions, presenting of the items). By adjusting the requirements of a CBT to a PPT both artifacts should be eliminated and equivalence should be established. Consistent with this suggestion, Neuman and Baydoun (1998) demonstrated that the differences across media can be minimized for clerical speed tests if CBT follows the same administration and response procedures as PPT. The authors concluded that their tests administered on different media measure the same construct but with different reliability.

Kim (1999) presented a comprehensive meta-analysis featuring two substantial enhancements over Mead and Drasgow's earlier work: First, the sample of 51 studies including 226 effect sizes was more heterogeneous including studies on classroom tests and dissertations. Second, the authors corrected for within-study dependency in effect size estimation using a method recommended by Gleser and Olkin (1994), thus, avoiding both the overrepresentation of studies with many dependent measures and the inflation of false positive outcomes. According to Kim, in a global evaluation of equivalence between computer-based and paper-based measures no differences across test media could be found as long as the testing is not adaptive.

In the recent past, two more meta-analyses (Wang, Jiao, Young, Brooks, & Olson, 2007; 2008) for mathematics and English reading comprehension respectively for K-12 students cover the research results of the last 25 years. For mathematics 14 studies containing 44 independent data sets allowed a comparison of the scores from PPT and CBT measures. After excluding six data sets contributing strongly to deviance in effect size homogeneity the weighted mean effect size was statistically not different from zero. One moderator variable, the delivery algorithm (fixed vs. adaptive) used in computerized administration, contributed statistically significant to the prediction of the effect size, whereas all other moderator variables investigated (study design, grade level, sample size, type of test, Internet-based testing, and computer practice) had no salient influence. For English reading assessment the weighted mean effect size was also not statistically different from zero after excluding six from 42 datasets extracted from eleven primary studies in an attempt to eliminate effect size heterogeneity. In comparison to the meta-analysis in mathematics, the moderator variables differ: Four moderator variables (study design, sample size, computer delivery algorithm, and computer practice) affected the differences in reading comprehension scores between test media whereas three other postulated moderator variables (grade level, type of test, and Internet-based testing) had no statistically meaningful influence. Even though, on a mean level no differences between test media could be found for mathematics and reading comprehension, the postulation of differential moderator variables for both disciplines might indicate a capitalization on chance or the relevance of unconsidered moderators (e.g., year of the study). Obviously the small sample of studies in both domains limits the generalizability of the results.

Considering all evidence presented so far, the effects of the test medium on test performance are nil or small. However, meta-analyses on the equivalence of ability measures across test media have a conceptual flaw. In order to adequately assess the equivalence across media a comparison of mean scores (and dispersions) is insufficient. Let us explain this point in more detail.

Horkay, Bennett, Allen, and Kaplan (2005) compared the writing performance of two nationally representative samples in a recent National Assessment of Educational Progress (NAEP) study. One sample took the test on paper, the other sample worked on a computerized version. Albeit, means in both conditions were roughly the same, computer familiarity consisting of a) hands-on computer proficiency, b) extent of computer use, and c) computer use for writing added about 11% over paper writing score to the prediction of writing performance in the PC-condition. Thus, students with greater hands-on skill achieved higher PC-writing scores when holding constant their performance on a paper-and-pencil writing test. Importantly, this difference in the construct measured by both instantiations would have remained undetected if the evaluation was solely based on tests of mean differences. So how does an appropriate procedure to test for equivalence between paper-based and computer-based measures look like?

How to test for equivalence?

Let us begin with the distinction between within- and between-subjects-designs in the context of cross-media-equivalence. Within the former the same subjects work both on paper and computer, within the latter different groups of subjects work either on paper or on computer. In both cases a potential test medium effect cannot be established by comparing or analyzing mean differences of the manifest or the latent scores. Strategies often applied in literature are based on the implicit assumption that the sources of within- and cross-media variance are actually the same. However, this assumption has to be tested by analyzing the means, variances and the covariances of the data. In this sense the framework of confirmatory factor analysis (CFA) provides the best method for equivalence testing for the measurement models used predominantly in educational and psychological measurement. In CFA the communality of many manifest variables is explained through a smaller number of underlying latent factors. This is achieved by, first, reproducing the covariance structure of the observed variables with the postulated covariance of a theoretical driven model, and second, evaluating the fit of the variable-reduced model to the empirical data. In case of a within-subject-design, the logic of testing is to check whether or not an additional, test medium specific factor accounts for unexplained variance and affects model fit beneficially. In case of a between-subject-design between-group comparisons and within-group comparisons are possible (for a detailed discussion see Lubke, Dolan, Kelderman, & Mellenbergh, 2003) by using *multi-group confirmatory factor analysis* (MGCFA).

Imagine the following within-subject-scenario: Subjects are working on three reasoning tests covering the content domains verbal, numerical, and spatial. All three tests with independent items sets are delivered within subjects on three different media: paper-pencil, notebook-computer, and personal digital assistant (PDA). After completing the three tests on one medium subjects continue with the next medium. The items are not identical across media but are drawn by the same mechanism from a predefined item universe. Therefore the tests are parallel from random fluctuations. Sequence effects are averaged out by balancing the design resulting in six different sequences.

As mentioned before, in order to test for equivalence in this example of a within-subject-design, first of all, a theory-driven structural model has to be established. In our case this could be a model with three correlated content-specific-reasoning factors (verbal, numerical, spatial) or a single factor model, also labeled g-factor model (cp. Wilhelm, 2005). The crucial step in equivalence testing lies in the introduction of one or more additional so called nested test medium factors. Nested factors are latent factors that are additionally to other latent factors loading on manifest variables (also called indicators) that have a method in common. Nested factors are usually conceptualized as uncorrelated to other factors of the measurement model. Thus, the variance of a variable is decomposed into a content-specific, a method-specific, and an error part. By introducing nested factors construct-irrelevant variance in the covariance structure could be tapped. The litmus test whether two measures are equivalent across media is to check if the introduction of a nested method factor makes a difference regarding model fit. This difference in model fit can be assessed descriptively with established fit indices like the *root mean square error of approximation* (RMSEA) or the *comparative fit index* (CFI) and inferentially predominantly with a *Chi-square difference test* (Bollen, 1989).

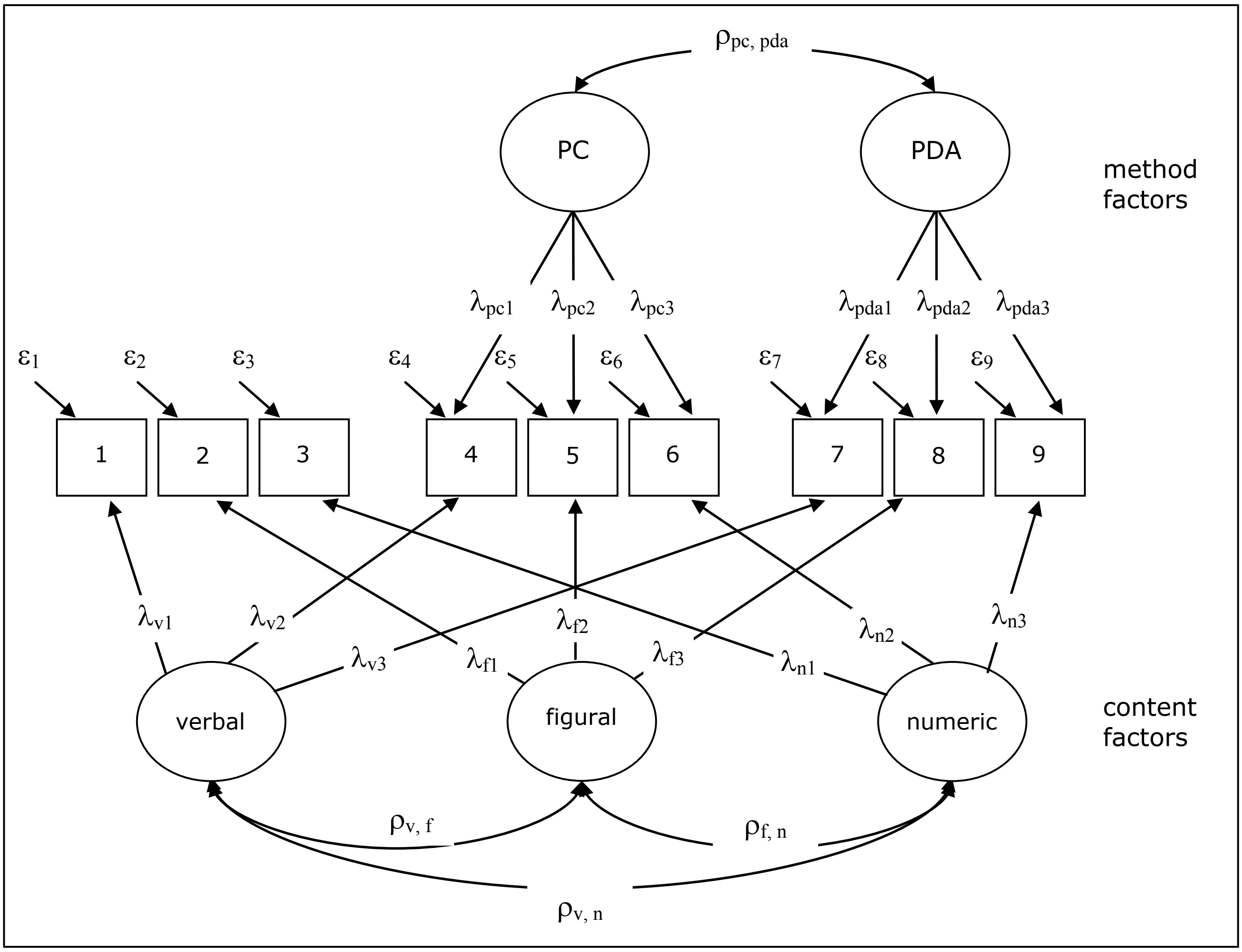


Figure 1: Correlated-trait-correlated-method-minus-one-model (CTC(M-1)-model)

In the multi-trait-multi-method-context different modeling strategies have been proposed to take method-specific variance – like the variance due to a test medium – into account. Depending on theoretical considerations a number of competing models for different purposes could be postulated. For instance, in case of inconsistent method artifacts on the indicators or an influence that is not unidimensional, correlated errors should substitute method factors, resulting in two possible models: *correlated-trait-correlated-uniqueness-model* (CTCU-model) and the *uncorrelated-trait-correlated-uniqueness-model* (UTCU-model, Widaman, 1985).

In the realm of equivalence testing one model seems exceptionally promising. In order to solve identification problems with the *correlated-trait-correlated-method-model* (CTCM-model) Eid (2000; Eid, Lischetzke, Nussbeck & Trierweiler, 2003) proposed the *correlated-trait-correlated-method-minus-one-model* (CTC(M-1)-model). Figure 1 depicts this CTC(M-1)-model with three traits and three methods in which one method is chosen as a reference. This reference is incorporated in the model by not specifying a separate method factor. As a consequence all specified method factors have to be interpreted in comparison to the reference method. In our example it would probably be sensible to choose the paper-pencil-condition as a reference method because we want to establish whether computer administration makes a difference in comparison to the traditional method of using paper and pencil.

On one side of the model in Figure 1 all content factors are correlated, on the other both method factors are correlated. The correlated method factors could be interpreted as a common computer-literacy-method factor that is orthogonal to the other factors in this model. An advantage of the CTC(M-1)-model is that the variance is totally decomposed into a trait-specific, a method-specific, and an error component enabling to relate them to each other. However, this benefit of uncorrelated content and method factors can also be seen as a disadvantage of this model architecture.

Because once method factors in the context of ability testing are interpreted it frequently turns out that those method factors might also express individual differences in methods and given the ubiquitous positive manifold amongst ability measures considering these method factors to be orthogonal to other ability factors is implausible. Nevertheless, in order to ascertain equivalence of data across media in a within-subject-design it is pivotal to check if the introduction of a method factor is improving model fit.

In the between-subject-design an extension of the CFA – *the multi-group confirmatory factor analysis* (MGCFA) – is a suitable method to check for equivalence of test data gathered with different test media. If you look on the (overarching) CFA approach in terms of a regression model the prediction of the observed score y for a person j is contingent on the latent ability level η. The prediction is based on an indicator *i* (e.g., an item of a general knowledge test) on a specific medium (e.g., PC). Formulized the relation is

$$y_{\mathrm{i,m,j}} = \tau_{\mathrm{i,m}} + \lambda_{\mathrm{i,m}} \cdot \eta_{\mathrm{m,j}} + \varepsilon_{\mathrm{i,m,j}}$$

where τ is the intercept, λ is the factor loading and ε is the residual term. To put it simple, in order to guarantee measurement invariance all these variables have to be equal across test media conditions. To understand the constituents of this formula more profoundly let us consider another example of a test measuring crystallized intelligence that was originally paper-based.

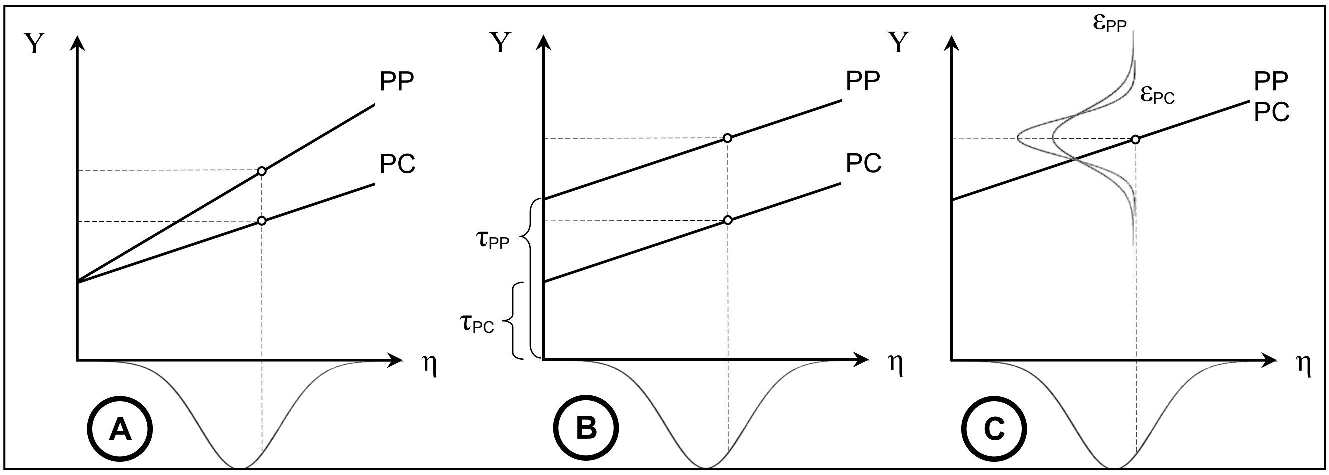


Figure 2: The consequence of divergent measurement parameters on the observed score. Note, that there is a perfect overlap between the ability distributions in both conditions (PP and PC).

This test is transferred to computerized delivery and the question of equivalence across test delivery methods has to be addressed. The three panels in Figure 2 describe various possible scenarios of measurement invariance for the crystallized intelligence test administered on both media, PP and PC. In the first panel (A) the subtests differ with regard to their slope or factor loadings ($\lambda_{\mathrm{i,PP}} > \lambda_{\mathrm{i,PC}}$). As you can see, the same ability level η results in different values of y. In the second panel (B) the difference lies in the intercepts ($\tau_{\mathrm{i,PP}} > \tau_{\mathrm{i,PC}}$). Here both functions run parallel, that is they have the same slope. Nevertheless, the same ability level η results in different values of y. With regards to content the difference between the intercepts amounts to the level of overprediction or underprediction, respectively. This situation of constant over- or underprediction independent of the ability level is referred to as *uniform bias* (Mellenbergh, 1982). In the third panel (C) the variance around the expected value is unequal implying different variances in the residual term ($\Theta_{\mathrm{i,PP}} \neq \Theta_{\mathrm{i,PC}}$). Even though the underlying ability distribution in both groups is the same, unequal model parameters cause differences in the observed scores. In other words, the different variances in the residual term produce measurement invariance or non-equivalence.

In order to ensure equivalence across test media all measurement parameters have to be equal in the regressions for all delivery methods. If there is a violation of these constraints there is some kind of invariance. Different levels of invariance across test media can be assessed with a straightforward procedure of comparing four models in a fixed order, from the least to the most restrictive model. A restriction is given when two measurement parameters are fixed to equality (e.g., $\lambda_{\mathrm{i,PP}} = \lambda_{\mathrm{i,PC}}$). Strong equivalence only holds if model comparison across the four consecutive steps are positive.

description	factor loading	residual variance	intercepts	factor means
symbol	Λ	Θ	τ	α
A configural invariance	Free	Free	Free	Fixed at 0
B metric invariance (weak factorial invariance)	Fixed	Free	Free	Fixed at 0
C residual variance invariance	Fixed	Fixed	Free	Fixed at 0
D strict factorial invariance (strong invariance)	Fixed	Fixed	Fixed	Free[1]

Table 2: Testing for equivalence in a between-subject-design with multi-group confirmatory factor analysis (MGCFA). Note. Free: Freely estimated within a group without any restrictions; Fixed: fixed to equality across groups; [1] The factor mean of one group is fixed at 0 whereas the factor mean of the other group is freely estimated.

Table 1 lists the different steps in invariance testing. In step 1 all measurement parameters (factor loadings, residual variances, and intercepts) are freely estimated in both conditions (PP and PC). Testing this stage checks for configural invariance. Here the pattern of loadings is more decisive than their actual magnitude. This is tested in step 2, metric invariance, where models are invariant with respect to their factor loadings whereas the other measurement parameters (residual variances and intercepts) are freely estimated. If measurement invariance is established on this stage, administration mode does not affect the rank order of individuals. This condition is also referred to as metric or weak invariance and is a prerequisite for meaningful cross-group comparisons (Bollen, 1989). In step 3, residual variance invariance, on top of the restrictions in step 2 the residual variances between groups are fixed to equality. In the most restrictive model (step 4) all measurement parameters are equal. If this standard is met strict factorial invariance (Meredith, 1993) holds. Wicherts (2007) explains why – in the last step in testing for strict equivalence – it is essential to allow for differences in factor means while fixing the intercepts to equality. Neglecting this possibility would force any factor mean difference in the group into differences in the intercepts, thus, concealing possible group differences. Each of the above models is nested within the previous ones, for example, model C derives from model B by imposing additional constraints. Due to this nested character a potential deterioration in model fit is testable through a *Chi-square-difference*-test. Cheung and Rensvold (2002) evaluated different goodness-of-fit indices with respect to a) their sensitivity to model misfit and b) dependency on model complexity and sample size. Based on a simulation they recommend using Δ(Gamma hat) and Δ(McDonald's noncentrality index) in addition to Δ(CFI) in order to evaluate measurement invariance. Although multi-group models are the method of choice in the between-subject scenario there are some interesting issues concerning: a) effect-sizes, b) location of invariance violation, c) application to ordinal measures, and d) the modeling of non-invariance (Millsap, 2005).

Discussion

In this chapter two methods have been presented that have a series of advantages over non-factorial approaches and clearly are more adequate than procedures frequently applied in the literature. In the discussion we want to focus on some heuristics on what can be done to achieve equivalence prior to collecting data. Because the testing is influenced by both software (e.g., layout design) and hardware aspects (e.g., display resolution) much effort has been devoted to answer this question from a technological perspective, for example, about the legibility of online texts depending on font characteristics (cp. Leeson, 2006). However, bearing in mind the rapid changes in soft- and hardware it seems hard to give long-lasting advice. Two simple heuristics that do not go far behind the obvious are: a) Only use technology that is essential to your measurement intention and b) use technology that is not restricted to specific machines or operating systems. Both recommendations are meant to avoid creating unnecessary prerequisites on hard- or software that restricts the potential sample and therefore affects internal validity (Cook & Campbell, 1979). From a psychological perspective chances are enhanced to establish equivalence across test media if the PC-condition is handled and as thoroughly scrutinized as a parallel paper-based test form. However, even a sound construction does not immunize against violations of stronger forms of equivalence. In this case it is inevitable and advisable to account for the additional source of variance. One way to accomplish this is to survey potential moderators like computer experience, accessibility to computers, and hands-on skills. As long as we do not know exactly why essential properties of ability measures vary across test media, investigating both equivalence and non-equivalence of computer- and paper-based test data is critical.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Bollen, K.A. (1989). Structural equations with latent variables. Oxford, England: John Wiley & Sons.

Cheung, G.W. & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling, 9, 233–255.

Cook, T. D. & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand-McNally.

Drasgow, F. & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. Journal of Applied Psychology, 70, 662-680.

Eid, M. (2000). A multitrait–multimethod model with minimal assumptions. Psychometrika, 65, 241–261.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. Psychological Methods, 8, 38-60.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. Journal of Educational Measurement, 39, 133-147.

Gleser, L.J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H.M. Cooper & L.V. Hedges (Eds.), The handbook of research synthesis (pp. 339-355). New York: Sage.

Goldberg, A.L. & Pedulla, J.J. (2002). Performance differences according to test mode and computer familiarity on a practice Graduate Record Exam. Educational & Psychological Measurement, 62, 1053-1067.

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: U.S. Department of Education, National Center for Education Statistics.

Kim, J.-P. (1999, October). Meta-analysis of equivalence of computerized and P&P tests on ability measures. Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chigaco, IL.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. International Journal of Testing, 6, 1-24.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. Intelligence, 31, 543-566.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychological Bulletin, 114, 449-458.

Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. Journal of Eduactional Statistics, 7, 105-118.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58, 525-543.

Millsap, R. (2005). Four Unresolved Problems in Studies of Factorial Invariance. Contemporary psychometrics: A festschrift for Roderick P. McDonald (pp. 153-171). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? Applied Psychological Measurement, 22, 71-83.

Overton, R. C., Taylor, L. R., Zickar, M. J., & Harms, H. J. (1996). The pen-based computer as an alternative platform for test administration. Personnel Psychology, 49, 455-464.

Wang, S., Jiao, H., Young, M.J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. Educational and Psychological Measurement, 67, 219-238.

Wang, S., Jiao, H., Young, M.J., 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, 5-24.

Wicherts, J.M., Dolan, C.V., & Hessen, D.J. (2005). Measurement invariance and group differences in intercepts in confirmatory factor analysis. Submitted. Available: http://www.test.uva.nl/wiki/images/b/b9/Measurement_invariance.pdf

Widaman, K.F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. Applied Psychological Measurement, 10, 1-22.

Wilhelm, O. (2005). Measuring reasoning ability (pp. 373-392). In O. Wilhelm & R.W. Engle (Eds.), Understanding and measuring intelligence. London: Sage.

The author:

Ulrich Schroeders
Institute for Educational Progress
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany
Tel.: ++49 (0)30 2093-9434

E-Mail: ulrich.schroeders@iqb.hu-berlin.de

Ulrich Schroeders received his diploma degree at the University of Wuerzburg and is currently a Ph.D. student at Humboldt-University in Berlin. His doctoral thesis addresses issues of ability test delivery (internet vs. lab; paper-pencil vs. notebook vs. PDA/Smartphone) for a broader collection of innovative measures.

Utilising the Potential of Computer-delivered Surveys in Assessing Scientific Literacy

Ron Martin
Australian Council for Educational Research (ACER), Australia

Abstract
This paper examines deficiencies in paper-and-pen-based science surveys and the potential for redressing these deficiencies with computer-based assessment. It looks at what was done in the PISA 2006 Computer-Based Assessment of Science optional component (CBAS) and how it addressed objectives that could not be assessed by paper-and-pen. Examples of interactive computer-based items that attempted to meet these objectives are discussed along with the methods for collecting relevant data arising from student interaction with the stimuli. The paper discusses reasons why the 2006 CBAS outcomes were not scaled with the paper-and-pen outcomes. Finally, the limitations on comparative international surveys arising from differences in 'computer cultures' between countries and regions are explored.

Educational policy decision makers may find sufficient reason for computer-based assessment of science in the future ubiquitous use of computing, and in the potential for cost-efficient assessment. Currently, any short-term movement toward this type of assessment is limited to those educational systems where there is universal provision of equipment of sufficient computing power in schools, security can be guaranteed, and there are cost savings to be made in coding. However, there are other reasons that can be attractive to educators and prove a strong motivation for change. The *Australian Council for Educational Research* (ACER), the successful contractor for the PISA 2006 Computer-Based Assessment of Science option (CBAS), made the following statement:

In focusing efforts to improve stimulus presentation, the Consortium will aim to enhance the array of stimulus types, to cut reading load, and to better target the vision of scientific literacy as laid out in the framework. Computer-based testing will allow the stimulus material to more realistically portray the real world context through video and vivid animations. Such stimulus is more likely to engage the student and therefore elicit a response that more accurately reflects their ability. Dynamic stimulus would also provide the opportunity to cover more of the scientific framework as more complex material can be presented and in a shorter, less reading intensive, time. (Extract from the ACER Consortium response to the Call for Tender, Feb. 2004)

As a consequence, the test developers for the 2006 CBAS option were given several objectives.

- Firstly, the computer-based test was to assess the skills and knowledge outlined in the Framework. This included those assessed in the Main Study provided this could be done in a value-added way. In particular, items that equally well could have been presented as paper-and-pen items were to be avoided.
- Secondly, aspects of scientific literacy that could not be adequately assessed in a paper-and-pen test were to be attempted in the computer-based test. In particular, dynamic stimuli were to be used and where applicable, interaction between the test-taker and the stimulus was to take place.
- Thirdly, where possible visual stimuli were to be used to reduce both the amount of text and the difficulty of text students had to contend with.
- Fourthly, by avoiding the use of coders, cost to countries was to be minimised and differences arising from differential keyboard skills were to be controlled. This meant there were to be no open-constructed response items other than those requiring only the use of a mouse. Such responses were to be captured by the software.

The computer-based test was seen as having the potential to replace paper-and-pen testing at some point in the future. Consequently, there was interest in scaling the data from CBAS with data from the Main Study paper-and-pen test.

Given the test development objectives it is clear that deficiencies in the paper-and-pen test were recognised and that an attempt was to be made to address them through the computer-based assessment. These deficiencies are summarised as an inability to

- describe complex contexts without generating a high reading load for students.
- convey dynamic contexts in which motion was an important factor.
- simulate investigations to show planning and measurement skills.
- allow for student intervention in simulated investigations to vary outcomes.
- assess the strategies students used in seeking information or evidence and to assemble records of data.

From the perspective of the test developers the computer-based test also offered the ability to determine the amount of time students needed to spend on each of the trialled items and consequently provided opportunities to better construct tests of appropriate lengths. It was also thought that students could better engage with the context if it were presented in a more dynamic way. The responses to attitudinal questions by students taking the computer-based option were to be a measure assessing the degree of engagement (CBAS Field Trial).

Achievements and failures of CBAS 2006

1. The following graphs indicate that a good match of competencies and knowledge types was achieved with the paper-and-pen test in the construction of the CBAS test. When the minor divisions are compared this match is not quite so good but given the small number of items in the CBAS assessment there is still a satisfactory match.

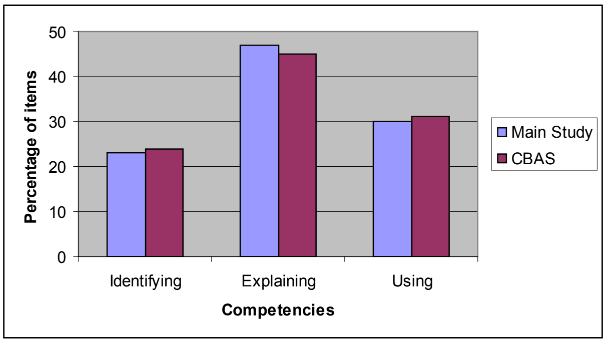


Figure 1: Comparison of Framework Classifications – CBAS and Main Study Competencies

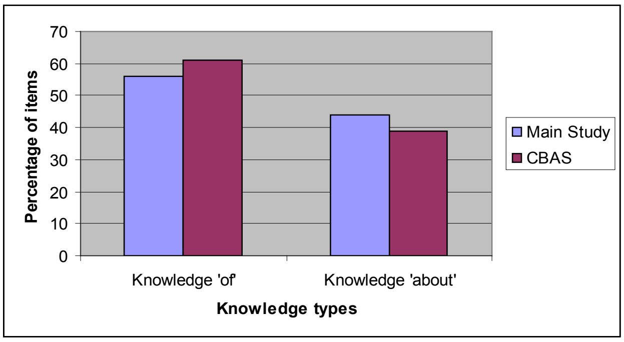


Figure 2: Comparison of Framework Classifications – CBAS and Main Study Knowledge Types

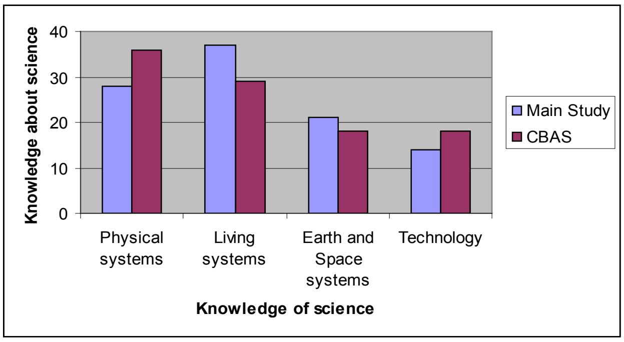


Figure 3: Comparison of Knowledge of Science Classifications

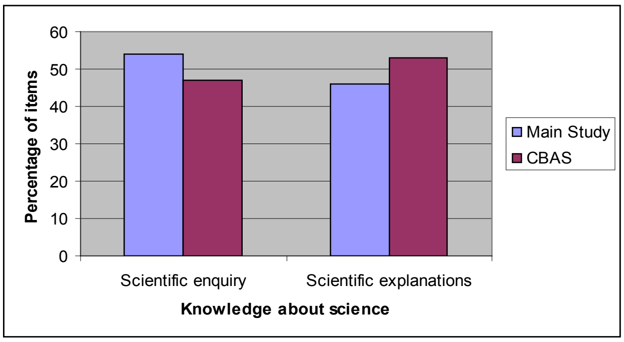


Figure 4: Comparison of Knowledge about Science Classifications

2. As a very rough gauge of readability, two measures have been compared – word count and the grade level. While taking into account any dot pointed phrases and incomplete sentences, the author has applied the *Flesch-Kincaid* software in Microsoft Word to the text in comparing the paper-and-pen and computer-based assessments for an estimate of grade level. This comparison has been restricted to similar items of a multiple-choice type. It needs to be emphasised that the grade level values given are not necessarily correct. Because of the small segments of text to which the method is applied it is likely that the grade level is overestimated. However, it is the comparison that is important and the same methodology is applied to both CBAS and the paper-and-pen tests.

Because the software tends to average the readability level over selected text the text in each item has been divided into stimulus, stem and options. In the case of the computer-based test this was rather straightforward as each page constituted stimulus, stem and options. However, in the case of the paper-based test one stimulus may provide the information for several questions. In this case to try and equate text readability, the number of words in the stimulus has been divided by the number of questions it 'services' and then added to each item's word count. If the grade level for the general stimulus was higher than for that of the stem or options in an item then that higher value alone was allotted to each of the items in the unit. In the case of the multiple choice options it was the option with the highest grade level readability score that was used for the options in that item as a whole.

The average number of words to be read per item in the CBAS test was 73.4. A similar methodology was used for the paper-and-pen test. The average number of words to be read per item in this form of the PISA science test was 105.7. Thus a reduction of about 30% was made in the amount of text to be read.

The correlation between the main study reading score and main study science score compared to the correlation between the reading score of students doing the CBAS option and the CBAS score supports the view that there was a considerable reduction in readability level in the CBAS test compared to the paper-and-pen test (Table 1).

	Reading
	r
Science Main Study Paper-and-pen test	0.84
Science CBAS test	0.75

Table 1: Latent correlations between reading and science (2006)

As can be seen in Figure 5, the Flesch-Kincaid grade level distribution is appreciably lower for the CBAS items than for the paper-and-pen ones.

Figure 5: Numbers of items in grade reading levels for CBAS and paper-and-pen tests – PISA 2006

3. Coding costs were eliminated by capturing the data by digital means. This created some problems in attempting to make the CBAS test of a form where responses would scale with the paper-and-pen responses. More is said of this later. Simulations and flash animations added to costs. However, there were numbers of video clips used which tended to contain costs except where new material had to be produced involving the use of film crews.

4a. The opportunity to simulate new conditions in investigations or change contexts opens possibilities in the assessment of features of scientific literacy not available in the paper-and-pen test. By altering variables students are able to control outcomes and apply their scientific knowledge in a comparative environment. The way in which students manipulate the conditions in simulating an investigation is of particular interest to science educators. A history of the interventions or responses students make in such simulations can be captured by the software and thus illustrate the strategies students have employed in carrying out that investigation. For example, in the units on fish farming and nuclear power plants the strategies students used could be recorded. Students who used methodologies where one variable was controlled while manipulating the other prior to seeking an optimum outcome for both can be distinguished from other students using random strategies where both variables were simultaneously varied. This data was not used in the PISA 2006 analysis but the potential for gaining greater information of student knowledge 'about' science is significant.

4b. Video clips were used to good effect in providing opportunities for students to apply their scientific knowledge in situations that could not be duplicated with a paper-and-pen form. Examples from the CBAS items include the movement of parts of an animal's body during its activities. Careful observation of this movement was essential for answering the question associated with this stimulus. In another unit the movement of two buildings during an earthquake tremor form the basis for student scientific explanation.

4c. What students will do in the planning and conducting of an investigation, without cues from the teacher, is usually assessed in practical hands-on contexts. In international surveys such as PISA and in some national surveys the time required and the use of similar equipment makes this impractical. Computer-based assessment offers some means of addressing this issue, mainly with simulations. No examples of trialled or piloted items like this survived to the PISA 2006 main study. However, an example can be found in some ACER simulations created to demonstrate ways of addressing this issue e.g. in a breeding experiment students have the option to test the outcomes of various cross-breeds of plants. Students make their own choices on how to proceed. The opportunity to replicate plantings is available and the means to record what students do with that opportunity.

4d. Dynamic illustrations of a phenomenon can be more inclusive than stimulus that relies on an ability to interpret through relatively abstract graphics. Examples of this form can be found in a CBAS unit where different methods of stopping a bicycle are dynamically illustrated, or in another unit where the effect on dolphins of a sonar warning signal attached to a net is examined. This latter example is a simulated experimental situation where the outcomes are automatically shown on a column graph that the students interpret. Items that incorporate actual observational measurement are also possible although there were no examples of this in the 2006 CBAS option.

5. Scaling: Given the nature of the objectives for the computer-based assessment it is surprising that there was any expectation that the test would scale with the paper-and-pen test. There were clearly differences built into the computer-based test that made it different from the paper-and-pen one. For instance, the text used was to be simpler and reduced in volume, there were to be no open-constructed response items of the form used in the paper-and-pen assessment, and aspects of scientific literacy that could not be properly assessed with paper-and-pen were to be included. The two types of assessment were analysed and the inappropriateness of scaling them together was demonstrated.

The results of that analysis from the trial in 13 countries are shown in tables II and III. It was evident in a comparison between a one-dimension and two-dimension modelling of the CBAS assessment (CBAS-c) and the paper-and-pen test done by students doing CBAS (CBAS-p) that a two–dimensional model fitted the data best.

- There is a significant difference ($p<0.0001$) between the final deviances for the two models [Table 2].
- The variance of the CBAS-c test is much smaller than that for the CBAS-p test [Table 2].
- The item fit statistics show the data fits a two-dimensional model better than a one-dimensional one [Table 3]
- Correlation between CBAS-c and CBAS-p can be considered moderate (Table 2). For instance, the correlation is no better than that between maths and science [Table 4].

	Unidimensional		Two-dimensional						Comparison		
Country	Final Deviance (1)	Reliability	Final Deviance (2)	Reliability of CBAS-p	Reliability of CBAS-c	Correlation Between CBAS-c and CBAS-p Items	Variance of CBAS-p	Variance of CBAS-c	ChiSq dif	df	p
1	39434	0.90	39402	0.881	0.869	0.931	0.784	0.577	32	2	0.0001
2	40644	0.85	40617	0.834	0.859	0.907	0.574	0.405	27	2	0.0001
3	25627	0.90	25574	0.845	0.889	0.918	0.861	0.461	54	2	0.0001
4	42871	0.85	42830	0.844	0.795	0.869	0.602	0.457	42	2	0.0001
5	25107	0.91	25068	0.920	0.891	0.890	0.785	0.560	39	2	0.0001
6	29174	0.86	29125	0.857	0.841	0.920	0.693	0.362	49	2	0.0001
7	32280	0.88	32216	0.853	0.822	0.863	0.645	0.348	64	2	0.0001
8	31570	0.87	31530	0.684	0.790	0.867	0.769	0.369	40	2	0.0001
9	42672	0.87	42624	0.833	0.850	0.906	0.658	0.415	48	2	0.0001
10	45633	0.89	45579	0.887	0.874	0.925	0.855	0.518	55	2	0.0001
11	32206	0.84	32181	0.782	0.779	0.892	0.551	0.371	25	2	0.0001
12	22689	0.71	22664	0.706	0.840	0.856	0.787	0.400	25	2	0.0001
Int.	430392	0.87	429963	0.835	0.839	0.901	0.683	0.466	429	2	0

Table 2: Comparison between one-dimension model and two-dimension model for CBAS-c and CBAS-p Items – [Field Trials]; [CBAS 2006 Preliminary Field Trial Data Analysis: Doc. CBAS(0510)2; Oct. 2005]

	Unidimensional model			Two-dimensional model		
	MNSQ	CI	T	MNSQ	CI	T
64 CBAS-	1.331	(0.954,1.046)	12.855	0.992	(0.954,1.046)	-0.328
116 CBAS-c	1.083	(0.956,1.044)	3.583	1.006	(0.956,1.044)	0.274

Table 3: CBAS-c and CBAS-p FIT Statistics); CBAS 2006 Preliminary Field Trial Data Analysis: Doc. CBAS(0510)2]; Oct. 2005

	Science	
	R	SE
Mathematics	0.89	0.0006

Table 4: Latent correlation between Science and Mathematics (Main Study – OECD)

Engagement with the contexts being assessed

One of the predicted advantages of using a computer-based assessment method was that students would find this more interesting, realistic and engaging than the paper-and-pen based test. Efforts were made to measure these effects through questions attached to the computer-based tests. The outcome was reported to a National Project Managers meeting in Australia in October 2005. It does appear that in those countries that participated in the trial, students had a preference for the computer-based assessment and this can be seen as one measure of engagement. There were differences in the way males and females expressed their preferences (Figure 6).

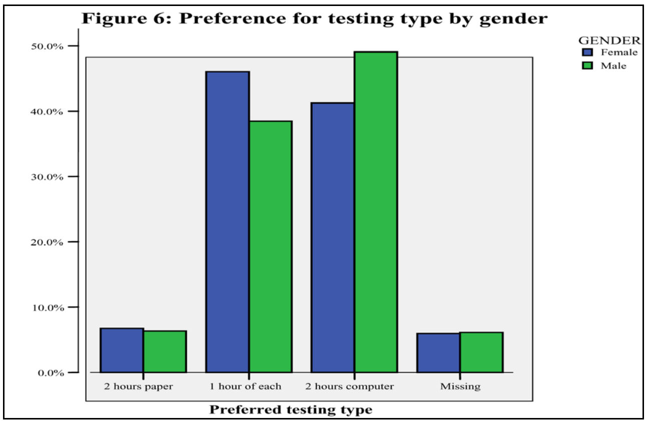


Figure 6: Preference for testing type by gender [CBAS 2006 Prel. Field Trial Data Analysis: Doc. CBAS(0510)2; Oct. 2005]

Limitations and projections

There are considerable differences in the readiness of students, both between countries and within countries, to use a computer-based form for assessing scientific literacy. Evidence from the field trial was that information communication technology (ICT) familiarity has a positive correlation with CBAS scores. This was true for both boys and girls although boys generally reported higher levels of ICT familiarity than girls (Figure 7).

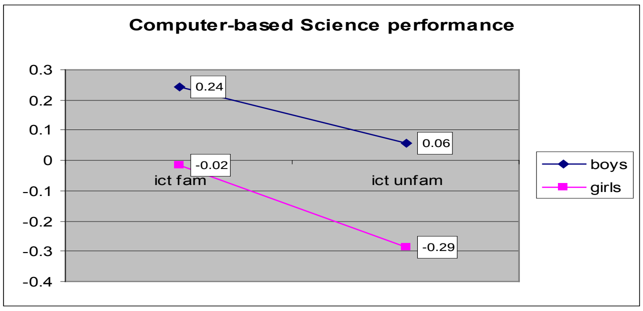


Figure 7: The effects of gender and ICT familiarity on CBAS-c performance - logits [CBAS 2006 Preliminary Field Trial Data Analysis: Doc. CBAS(0510)2; Oct. 2005]

Since this familiarity is likely to differ markedly between countries it is unlikely that a computer-based assessment would become the preferred method for assessing scientific literacy for international surveys such as PISA and TIMSS for a couple of cycles to come. However, in an increasingly IT literate world, the potential of computer-based assessment to more broadly assess the objectives of scientific literacy education is high. It is likely that some countries, with modern and uniform distribution of IT facilities, able to put considerable resources into the development of dynamic and interactive items and a nationwide teaching staff well-trained in the use of IT and actively using it in their classrooms, will be able to move into this sort of assessment relatively soon.

References

Australian Council for Educational Research (2007) Demonstration computer-based science units, Internal files

Consortium (2005) CBAS 2006 Preliminary Field Trial Analysis, Doc: CBAS(0510)2; Internal document.

Consortium (2006) PISA Database

Consortium (2006) CBAS Main Study Forms 1 and 2

Consortium (2004) Response to Call for Tender; Internal document.

Consortium (2005) CBAS Trial Clusters

Davidson, A. & Green, G.M. (eds) (1988) Linguistic complexity and text comprehension: readability issues reconsidered; Hillsdale, N.J., L. Erlbaum Associates.

Farr, J. N., Jenkins, J. J., and Paterson, D. G. (1951), Simplification of Flesch Reading Ease Formula, Journal of Applied Psychology, Volume 35, Number 5, (October), pp. 333-337

Flesch, R.F. (1974) The art of readable writing with the Flesch readability formulas; Harper & Row, New York

OECD (2006) Assessing scientific, reading and mathematical literacy: A framework for PISA 2006, OECD publishing, Paris.

OECD (2007) PISA 2006: Science competencies for tomorrow's world, Volumes 1: Analysis, and Volume 2: Data, OECD publishing, Paris.

The author:

Ron Martin
Australian Council for Educational Research (ACER)
19 Prospect Hill Road
Camberwell, Vic. Australia 3124
E-Mail: martin@acer.edu.au

Dr. Martin is a Senior Research Fellow with the Australian Council for Educational Research. His work involves assessing scientific literacy across the primary and secondary education sectors in an Australian and worldwide context. He was part of the science development team for the Programme for International Student Assessment (PISA) and the computer-based optional component of that programme. His research interests are in developing assessment tools to cover all aspects of science education required for the citizen of today's societies and those with the potential for careers as science professionals.

Are Icelandic boys really better on computerized tests than conventional ones?

Interaction between gender, test modality and test performance

Almar M. Halldórsson, Pippa McKelvie & Júlíus K. Björnsson
Educational Testing Institute, Iceland

Abstract:
Iceland has participated in OECD´s Programme for International Student Assessment (PISA) since the first study in 2000. In PISA 2003 Iceland was the country where girls had the greatest advantage over boys in reading literacy as well as in mathematics. The PISA 2006 cycle included an optional computer-based component assessing scientific competencies (Computer-Based Assessment of Scientific Literacy - CBAS) and Iceland's participation in CBAS was intended to investigate this gender gap finding. This article examines modality effects on gender performance by comparing achievement results on the PISA 2006 paper-and-pencil (P&P) assessment and CBAS. Gender difference is compared in terms of several factors relating to both student aptitude and item specific factors. These include: Computer familiarity, motivation, enjoyment, effort on the test, interactivity of computer items, reading load of items and item difficulty. A clear-cut finding is that boys in all three participating countries (Iceland, Denmark and Korea) outperformed girls in science literacy when the test was presented via computer regardless of the patterns of achievement across gender on the PISA paper-and-pencil test. Despite the intuitive relationship between higher motivation, greater experience with and confidence for ICT tasks and achievement on the computer-based test, statistical analysis of the correlations between achievement and these factors did not reveal any significant association with achievement. The increase in boys' performance in CBAS may however be partially explained by lower reading load and by boys' greater test fatigue on low difficulty items in paper based tests. Gender difference favouring girls in Iceland is removed in performance on paper based items of low reading load (under 100 words) so it is proposed that the difficulty of the P&P science items may fatigue boys and encourages them to 'give up' on P&P tests more than girls. Boys may be disadvantaged by the length of the P&P science items. Some cautionary notes are made about further studies with balanced test design, similar experiments should use a third reference group where a group of matched students are given the same paper-and-pencil items via computer.

Iceland has participated in *the Programme for International Student Assessment* (PISA) since the first study in 2000. The PISA study has shown a strong female advantage in Iceland for students age 15, compared to other countries. The results for 2000 indicated that the gender gap in reading literacy favouring girls was substantial in Iceland. However, no gender difference was found in mathematics and science literacy that year. In PISA 2003, Iceland was noted as the only country where girls performed significantly better than boys in mathematics. In all 41 participating countries, Iceland was also the country where girls had the greatest advantage over boys in reading literacy. Furthermore, Iceland was one of only three countries where the science literacy of girls was higher than for boys. In PISA 2003 a special test was administered to assess problem solving skills of students and the greatest gender difference favouring girls by far was found in Iceland.

The PISA results on gender differences in Iceland received international media attention [1] and spurred further research in Iceland where gender difference in educational achievement had already been the subject of extensive research (see, for example, Jóhannesson, 2004; Magnúsdóttir, 2006; Ólafsson et al., 2006 and Ólafsson et al., 2007). There are indications of gender specific learning cultures, where learning plays a very different role for girls than for boys in the socialisation processes in adolescence. Magnúsdóttir's (2006) research indicates that getting high marks is part of the image of a girls' leader, while for boys' leaders high marks are not as important. Research by Kristjánsson et al. (2005) shows that a higher proportion of girls believes it is important to do well at school. More girls claim they intend to study at university and they like school more than do boys. Sigfúsdóttir (2005) also shows that the "cultural capital" of girls is greater than boys': They get greater support from their families, they are more often required to follow rules than boys, parents know their friends better, etc.

The PISA survey is administered in the standard paper-and-pencil format. More than 400 000 students from 57 countries took part in the PISA 2006 assessment. The focus in this assessment cycle was on science literacy and the assessment also included an optional computer-based component assessing scientific competencies (*Computer-Based Assessment of Scientific Literacy* - CBAS). Three countries administered the CBAS component (Denmark, Iceland and Korea).

Previous studies have indicated that use of computers in the home (and greater ICT confidence) is strongly correlated with higher academic achievement (Harrison et al., 2003; Ravitz et al., 2002). Further research specifies that only home use of computers for educational purposes was associated with higher performance (in mathematics), whereas out-of-school use of ICT was negatively associated with performance (Valentine et al., 2005).

Notably, computer-based assessment requires fewer language skills, can present more information succinctly and in a shorter space of time. It is particularly useful in the assessment of science for simulating scientific phenomena that cannot easily be observed in real time such as seeing things in slow-motion or speeded-up, for modelling scientific phenomena that are invisible to the naked eye (e.g., the movement of molecules in a gas). This presents students with the opportunity to perform repeat trials in limited testing time, or for working safely in lab-like simulations that would otherwise be hazardous or messy in a testing situation.

Iceland’s participation in CBAS was in a substantial way based on the large gender gap finding in previous cycles of PISA. One hypothesis states that boys could potentially outperformed girls on computer-based items because they are more competent in and familiar with the types of ICT tasks required of them to complete the items due to their greater ICT familiarity. However, OECD’s PISA 2003 ICT report revealed that greater Internet use and program use were actually associated with a drop in mathematics and reading performance, stating that “one cannot readily assume that computer usage is bound to be beneficial for students in all cases” (OECD, 2005b, p.65).

An important question is how much of gender difference in test performance can be contributed to the modality of the test, the way the material is presented and student´s engagement in the test situation. This article examines modality effects on gender performance by comparing achievement results on the PISA 2006 paper-and-pencil assessment of science with performance in the CBAS 2006 computer-based component. Gender difference is compared in terms of several different factors relating to both student aptitude and item specific factors. These include: computer familiarity, motivation, enjoyment, effort on the test, interactivity of computer items, reading load of items and item difficulty.

Method

Sample

All students participating in PISA and CBAS in 2006 were born in 1990. A subsample of 100 schools was selected to participate in CBAS from the main PISA 2006 school sample test in Iceland. From these schools, clusters of 5 to 45 PISA-eligible students were sampled from the PISA student sample. All schools and students selected for CBAS had already participated in the paper-and-pencil PISA 2006 assessment.

It is important to note that the sample considered in the present analyses includes all students that participated in the CBAS test session as well as all PISA-participating students from the schools that had at least one student participating in CBAS. For Iceland, the original CBAS sample was drawn with 1104 students out of which 784 students participated (71% response rate). However these analyses include data for an additional 2782 students who participated in the paper-and-pencil assessment of science, attended a CBAS-participating school but did not respond to the CBAS test. To give achievement scores on the CBAS test for these students, plausible values on the CBAS scale were statistically imputed based on the students’ PISA paper-and-pencil achievement and background information. A total of 3566 students are therefore included in the CBAS analyses for Iceland, which is very close to the total number of students participating in the paper-and-pencil PISA 2006 (3789). As a result, we can be confident that the Icelandic sample for CBAS is representative of the population of 15 year old students in the country.

To account for any biases in selection of schools and students, the PISA data are weighted using a balanced repeated replication method. This accounts for, for example, any over- or under- representation of geographical areas within countries. More information about the weighting techniques in PISA can be found in the Data Analysis Manual (OECD, 2005a) or in the PISA 2006 Technical report (OECD, forthcoming).

In Denmark and Iceland the CBAS sample was approximately equally constituted of boys and girls but in Korea there is a greater number of boys than girls (see Table 1).

	CBAS sample	
Country	*Girls*	*Boys*
Denmark	52%	48%
Iceland	50%	50%
Korea	44%	56%

Table 1: Proportions of girls and boys in the sample analysed in this report

Procedure

CBAS test sessions took place either on the same day as the PISA paper-and-pencil assessment of students' reading, mathematics and science performance, or very shortly thereafter. Test administration was standardised so that all students performed the test on the same type of laptop, using the same software and in a similar testing environment. Up to five students participated in each test session under the guidance of one Test Administrator. The computer-based science items were presented to students on laptop computers through CBAS software specially designed for this purpose. This was a fixed-form test where the same 45 items were presented to all students in one of either two orders. The order of items was split from the middle point of the second form so as to reduce fatigue effects on the items occurring later in the test.

The software allowed students to move between items as they wished and to return to questions (changing their answers if necessary) up until 55 minutes had elapsed since the beginning of the test, at which point the Test Administrator stopped the session. This allowed for just over 1 minute per question. If a student finished early the items remained on the screen until the completion of the 55 minute test session. Following the cognitive items questionnaire items were presented and students had 5 minutes to respond to these. In total therefore, test sessions were one hour long.

Hardware

All laptops used for student testing were required to comply with a number of minimum specifications: A CPU 1.6 GHz Pentium M Processor, memory 512 of RAM, hard disk 40 GB, display 14.1" XGA, an optical mouse, external stereo headphones and the operating system Windows XP Professional.

Cognitive items

In total, 45 items with multimedia extensions (animations using flash software, video footage or photos) were presented to students. The final analyses are performed on 42 items as two items were dropped prior to the analyses and two items were combined into one as they were considered to be assessing the same knowledge. Two additional items were set to 'not administered' for Icelandic students, one showed video footage of a vitamin tablet dissolving in water which was judged as an unfamiliar concept for Icelandic students and in one of them specific terms in the item were not translatable into Icelandic. All item designs were either multiple choice or complex multiple choice involving, for example, a number of Yes/No responses for the answers offered. A small number of complex multiple choice items asked the students to place items in a specific order or position in a given diagram.

Scaling

Initially, CBAS scores for the three countries (Iceland, Denmark and Korea) were scaled on the traditional PISA scale with a mean of 500 and a pooled SD of 100. Paper-and-pencil (P&P) Science, Reading and Mathematics scores for the three CBAS countries were also re-scaled from the same model as the CBAS plausible values so as to allow calculation of correlations between CBAS and the paper domains. Because these new scores were re-scaled for only 3 countries, they are not directly comparable with the OECD-reported PISA 2006 test scores where 500 and 100 are the mean and SD of all OECD countries. Therefore, to avoid confusion between the scales all achievement scores have been re-standardised on a new scale with a mean of 5 and a SD of 2, meaning that over 99% of students have scores between 0 and 10 on the scale. This removes the possibility of direct comparisons between the scores reported here and the scores reported in the OECD PISA 2006 report which would not be valid because the plausible values are drawn from different models.

Moderating factors

Item difficulty: Item difficulties were calculated and Figure 1 below shows that the items were approximately evenly distributed across the item difficulty scale from -3 to 3 with the mean item difficulty at zero, indicating good coverage of all potential competency levels. Percentage correct per CBAS item was also calculated and ranged from 13% to 94% with an average percentage correct per item of 60%. Percentage correct per item was strongly associated with item difficulty from the model at 0.90 indicating that percentage correct per item is also an adequate measure of performance for specific analysis purposes.

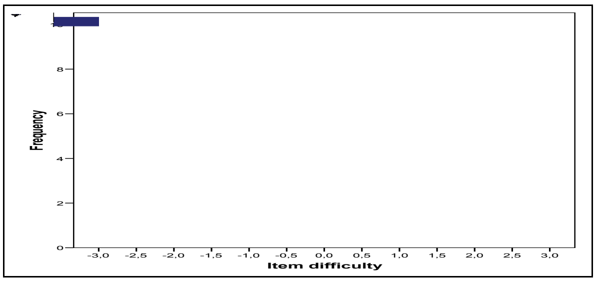
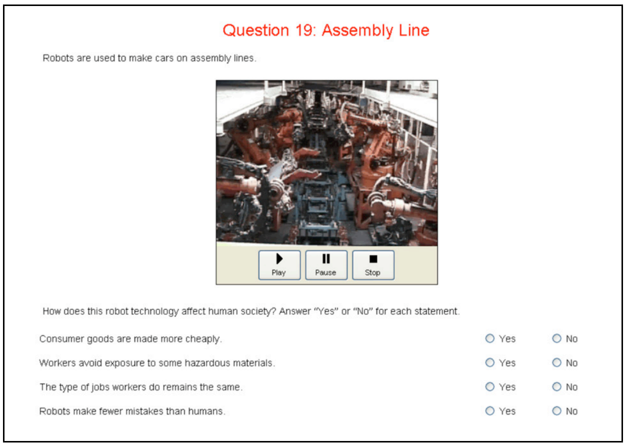


Figure 1. Distribution of Item difficulty for final CBAS items (Mean: -0,01027; StDev: 1,11964)

Interactivity: As the computer-based items differ markedly from the P&P items in terms of how much the student can interact with the item (for example, the possibility of moving levers to adjust levels in experimental trials or dragging and dropping the answer into the correct location in the diagram) an important effect across gender is interactivity of the items.

A panel of three independent judges rated all CBAS items into three groups according to the level of interactivity (low, medium and high) based on the types of activities the student had to perform with the item and based on how much the student needed to engage with the audiovisual material to answer the question. An example of a low interactivity item is the "Assembly Line" item in Figure 2 that shows a short video of an automated car assembly line and asks a question related to the role of robots in society.



Question 19: Assembly Line

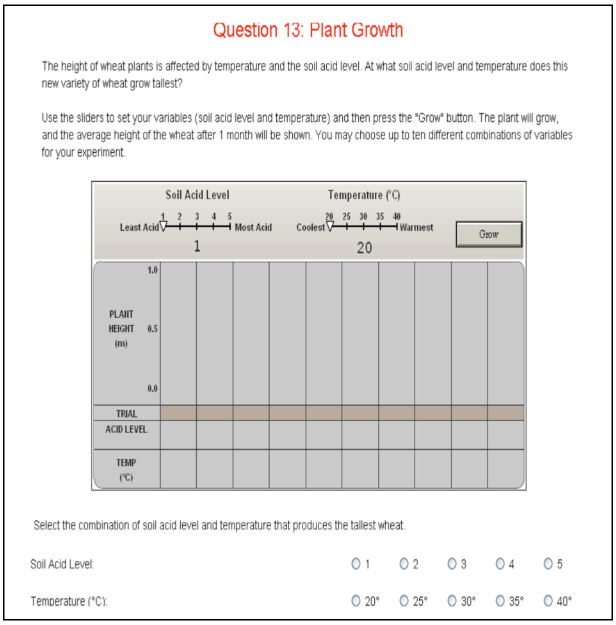
Robots are used to make cars on assembly lines.

How does this robot technology affect human society? Answer "Yes" or "No" for each statement.

Consumer goods are made more cheaply.	○ Yes	○ No
Workers avoid exposure to some hazardous materials.	○ Yes	○ No
The type of jobs workers do remains the same.	○ Yes	○ No
Robots make fewer mistakes than humans.	○ Yes	○ No

Figure 2. Sample unit: Assembly Line

Here, the video footage serves as contextual information to the item but does not provide the answer. In fact, this question could be answered correctly without the student watching the video footage and is therefore considered to be of low interactivity. In contrast, the following item in Figure 3, "Plant Growth", where the student is required to move buttons up and down a scale, performing experimental trials on optimal temperature and soil acidity levels for growing wheat, was considered as highly interactive.



Question 13: Plant Growth

The height of wheat plants is affected by temperature and the soil acid level. At what soil acid level and temperature does this new variety of wheat grow tallest?

Use the sliders to set your variables (soil acid level and temperature) and then press the "Grow" button. The plant will grow, and the average height of the wheat after 1 month will be shown. You may choose up to ten different combinations of variables for your experiment.

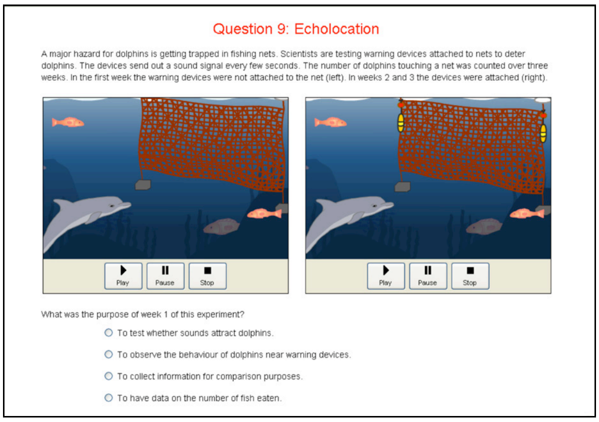
Select the combination of soil acid level and temperature that produces the tallest wheat.

Soil Acid Level:	○ 1	○ 2	○ 3	○ 4	○ 5
Temperature (°C):	○ 20°	○ 25°	○ 30°	○ 35°	○ 40°

Figure 3: Sample unit of highly interactive item: Plant Growth

Overall, fourteen items were classified as high interactivity, thirteen as medium and sixteen as involving low interactivity.

Reading load: Word counts for each CBAS item were recorded including the number of words in the stimulus, embedded in the image, in the question stem and in the multiple choice response options. Based on these figures, the CBAS items were divided into three groups according to reading load: low, medium and high. Eleven items were considered to be of a high reading load, for example as shown in Figure 4:



Question 9: Echolocation

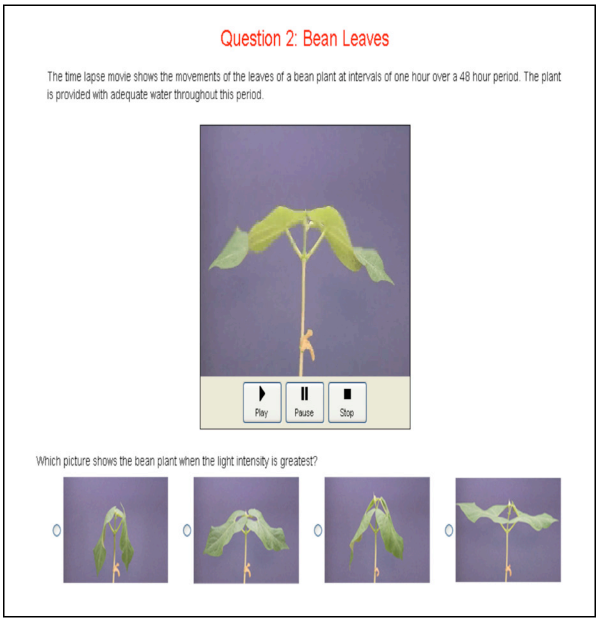
A major hazard for dolphins is getting trapped in fishing nets. Scientists are testing warning devices attached to nets to deter dolphins. The devices send out a sound signal every few seconds. The number of dolphins touching a net was counted over three weeks. In the first week the warning devices were not attached to the net (left). In weeks 2 and 3 the devices were attached (right).

What was the purpose of week 1 of this experiment?

○ To test whether sounds attract dolphins.
○ To observe the behaviour of dolphins near warning devices.
○ To collect information for comparison purposes.
○ To have data on the number of fish eaten.

Figure 4: Sample item showing high reading load item: Echolocation (Q3)

Fourteen items were classified as medium reading load and eighteen items (including the item in Figure 5) were classified as low reading load.



Question 2: Bean Leaves

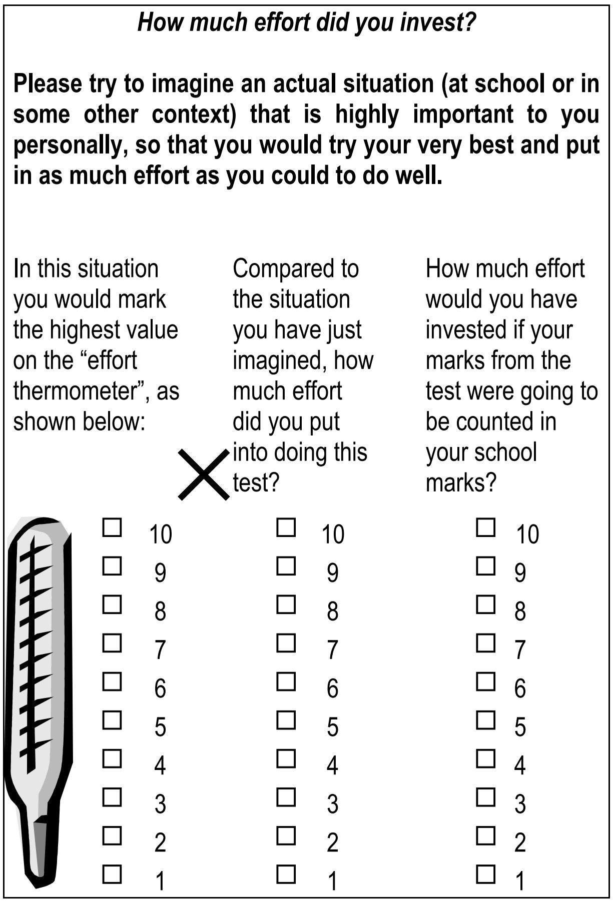
The time lapse movie shows the movements of the leaves of a bean plant at intervals of one hour over a 48 hour period. The plant is provided with adequate water throughout this period.

Which picture shows the bean plant when the light intensity is greatest?

○ ○ ○ ○

Figure 5: Sample item showing low reading load item: Bean Leaves (Q1)

Motivation, Enjoyment and Effort: In CBAS, after the cognitive items the students were asked to respond to several short questions to investigate the effects of enjoyment, motivation and effort on performance. Students were asked to rate on a four-point Likert scale how much they enjoyed the computer-based and paper & pencil tests, and whether they would do a similar test where the answers were provided "just for fun" (assessing motivation). The PISA Effort Thermometer was also used where students were asked to imagine an actual situation that was highly important to them personally, so that they would try their very best and put as much effort as they could to do well. They were told that in this situation they would mark the highest value on the effort thermometer (10) and then they were asked to report: how much effort they put into doing the CBAS test compared to the situation they had just imagined; and how much effort they would have invested if their marks from CBAS had counted in their school marks. This questionnaire item was identical to the item used in the PISA paper-and-pencil test and is displayed in Figure 6:



How much effort did you invest?

Please try to imagine an actual situation (at school or in some other context) that is highly important to you personally, so that you would try your very best and put in as much effort as you could to do well.

In this situation you would mark the highest value on the "effort thermometer", as shown below:	Compared to the situation you have just imagined, how much effort did you put into doing this test?	How much effort would you have invested if your marks from the test were going to be counted in your school marks?
□ 10	□ 10	□ 10
□ 9	□ 9	□ 9
□ 8	□ 8	□ 8
□ 7	□ 7	□ 7
□ 6	□ 6	□ 6
□ 5	□ 5	□ 5
□ 4	□ 4	□ 4
□ 3	□ 3	□ 3
□ 2	□ 2	□ 2
□ 1	□ 1	□ 1

Figure 6: PISA Effort Thermometer

In addition, students were asked which test they put more effort into between the CBAS test and the PISA paper test (assessing relative effort) and what type of test they would prefer between a two hour paper-and-pencil test, one hour of each type of test and two hours of computer-based testing.

ICT Familiarity: All countries participating in CBAS also administered the PISA ICT questionnaire during the PISA paper-and-pencil questionnaire session (along with 37 other countries which contribute to the calculation of the scale indices). This questionnaire has 32 questions about the frequency of computer use for specific activities and confidence in performing specific activities on the computer. Two scale indices were computed from measuring ICT familiarity: Internet/entertainment use and Program/software use. More information can be found on these indices in the OECD report on PISA 2006 (OECD, 2007).

Table 2 shows the model fit for a four-dimensional model for the ICT familiarity items in PISA 2006. Fit indices measure the extent to which a model, based on a particular structure hypothesised by the researcher, 'fits the data'. Model fit is assessed using Root-Mean Square Error of Approximation (RMSEA), the Root Mean Square Residual (RMR), the Comparative Fit index (CFI) and the Non-normed Fit index (NNFI). The PISA 2006 Technical Report should be consulted for further information about these techniques (OECD, forthcoming). Overall, the model fit was considered satisfactory for all of the CBAS participating countries and for the pooled OECD sample.

	Model fit			
*Country**	*RMSEA*	*RMR*	*CFI*	*NNFI*
Denmark	0.099	0.084	0.69	0.70
Iceland	0.089	0.078	0.71	0.72
Korea	0.077	0.060	0.79	0.80
OECD	0.084	0.082	0.81	0.81

Table 2: Model fit for CFA with ICT familiarity items
* Model estimates based on international student calibration sample (500 students per OECD country).

Table 3 shows the scale reliabilities for the ICT scales in CBAS countries and the overall median for all PISA countries that administered the ICT familiarity questionnaire. The internal consistencies were mostly high across all PISA countries but are well below the median for all CBAS countries. These scales may therefore be slightly less reliable in the CBAS countries than in the PISA countries as a whole. These scales are nonetheless used here and considered to be a fairly good estimate of ICT familiarity in the CBAS countries.

Country	*Internet /entert. use*	*Program use*
Denmark	0.66	0.73
Iceland	0.69	0.75
Korea	0.66	0.71
Median	*0.82*	*0.78*

Table 3: Scale reliabilities for ICT familiarity scales

Results

Gender difference in performance on the PISA P&P test and the CBAS computer test in science literacy is considered in light of a number of moderator variables described above. First, gender differences in performance across test modalities in the three CBAS countries are considered, then findings are discussed in terms of ICT familiarity, motivation, enjoyment and effort. Finally, results on interactivity of computer items, reading load and item difficulty are reported.

Gender Differences in student performance across test modalities

As Figure 7 shows, in Denmark, boys performed significantly better than girls on the P&P test of science by almost ¼ of a standard deviation. In Iceland, girls slightly outperformed the boys on the P&P test of science and in Korea there were no significant gender differences. The gender differences were large and clearly directional when science achievement was tested via computer however, with boys performing better than girls on the CBAS test in all countries.

Boys outperformed girls on CBAS by approximately ¼ of a standard deviation in Iceland and Korea to almost half a standard deviation in Denmark. Denmark has the largest gender difference in favour of boys regardless of test modality, but it should be noted that the increase in size of the gender difference as students moved from one test to the other is similar across all three countries. In other words, in Denmark the gender advantage for boys increased by ¼ of a standard deviation from ¼ to ½. In Korea it also increased by ¼ of a standard deviation from 0 to ¼ and in Iceland the increase was slightly larger as the advantage was reversed from $^{1}/_{10}$ of a standard deviation to ¼ of a standard deviation.

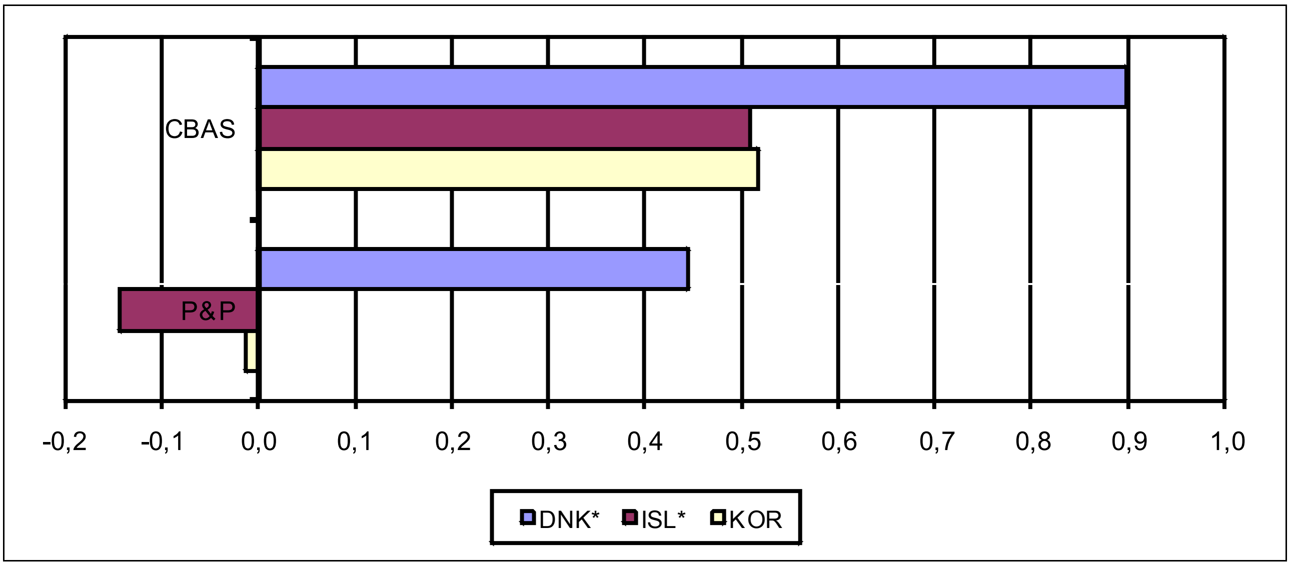


Figure 7: Boys' achievement advantage across tests and countries (positive values show boys outperforming girls)

When we compare the mean achievement scores for girls and boys in the paper-and-pencil test to the CBAS test across countries in Table 4 we can see that in Denmark boys' and girls' CBAS performance dropped (with girls' performance dropping more than boys increasing the gender difference). In Iceland and Korea, boys' performance increased, leaving behind the girls whose respective performance decreased and creating the gender difference seen earlier. Statistically significant differences are tested with the means and standard errors of the mean calculated through the replicates procedure involving the eighty PISA replicate weights on plausible values. When a gender difference is statistically significant at the $p<0.05$ level of significance, the boys and girls means have been printed in bold in the following table.

	Paper & Pencil			*CBAS*		
Country	*Girls*	*Boys*	*Total*	*Girls*	*Boys*	*Total*
Denmark	**4.39 (0.16)**	**4.85 (0.14)**	4.62 (0.12)	**3.81 (0.15)**	**4.71 (0.12)**	4.25 (0.11)
Iceland	**4.49 (0.05)**	**4.34 (0.05)**	4.41 (0.04)	**4.18 (0.04)**	**4.69 (0.05)**	4.44 (0.03)
Korea	5.06 (0.13)	5.03 (0.12)	5.04 (0.09)	**4.79 (0.14)**	**5.31 (0.13)**	5.08 (0.10)

Table 4. Achievement in Paper & Pencil test of Science compared to CBAS (se of mean)

The correlations in Table 5 further show that girls' CBAS performance is slightly less strongly associated with their performance on the P&P test of science than for boys, indicating that the impact of changing the test method is not the same for girls as it is for boys.

Correlations	*Girls*	*Boys*
Denmark	0.89	0.91
Iceland	0.78	0.80
Korea	0.88	0.90

Table 5: Correlations between P&P science scores and CBAS scores across genders and countries

Table 6 presents this relationship in another way, displaying the correlations for familiarity and achievement across countries which are stronger for boys than for girls (although on the whole they are relatively weak).

	P&P		*CBAS*	
	Girls	*Boys*	*Girls*	*Boys*
Denmark	0.03	0.09	0.06	0.11
Iceland	-0.04	0.07	-0.02	0.10
Korea	0.05	0.07	0.07	0.10

Table 6. Correlations between ICT familiarity science scores on the CBAS and P&P tests.

In the paper-and-pencil PISA 2006 results (OECD, 2007) it was reported that girls performed significantly better overall than boys on the Knowledge about Science items (which combine both the Scientific Explanation and the Scientific Enquiry items). This general pattern was also present in the Icelandic data shown in Figure 8 below using a percent correct calculation. Here we see that girls outperform boys on the items assessing the methods of science (Knowledge about Science), whereas overall boys have the advantage on the Knowledge of Science items (despite slight advantages for girls in the Living Systems and Earth & Space Systems questions).

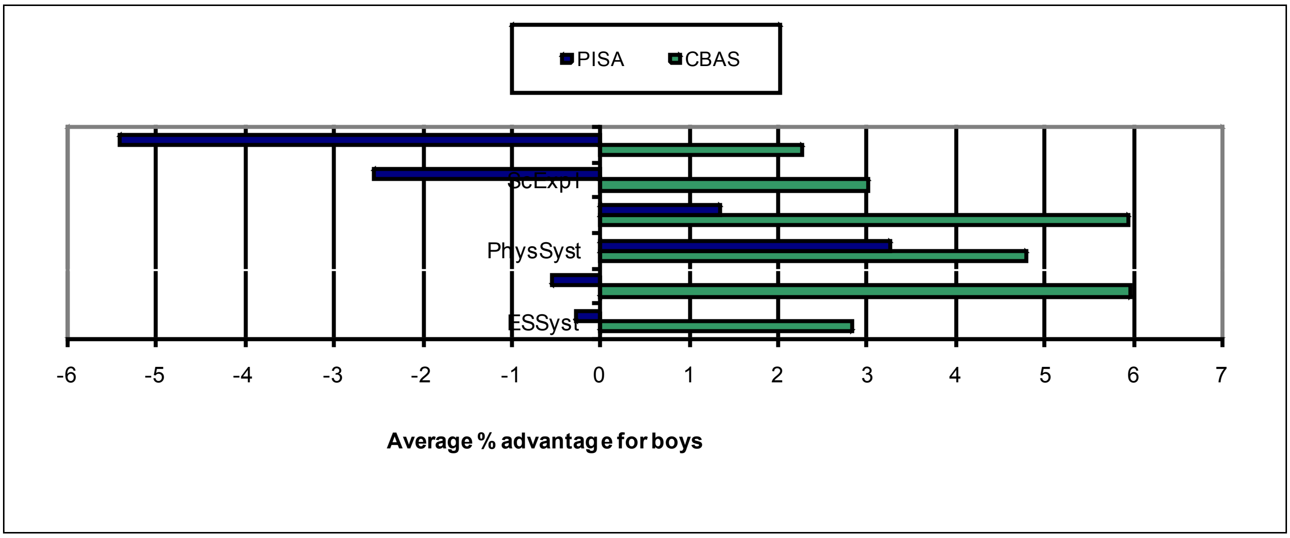


Figure 8:Average advantage in performance across domains for CBAS and P&P science items in Iceland

When we compare the performance in domains across test modalities, it is interesting to note that the boys' advantage decreases in the same domains that the girls displayed strengths in on the paper-and-pencil test (Knowledge about Science). This is a good indication that the items across both tests are assessing the same competencies as the gender difference changes in the same manner across domains regardless of test modality.

A general finding from these comparisons is that no country interaction is found for the relationship between science achievement on the paper-and-pencil test and science achievement on the computer-based test. There seems to be an overall effect in all three countries where boys outperform girls on the computer-based assessment of science in all countries irrespective of the gender difference in the paper based test.

Familiarity with Information Communication Technologies (ICT) across genders and impact on achievement

The following analyses investigate the relationship between the measure of ICT familiarity and boys' and girls' performance in CBAS and P&P science for each country. Overall, boys score higher on the frequency of use scales than girls and more ICT familiar students perform better in CBAS than less ICT familiar students irrespective of age, although results for Icelandic girls are an exception.

In Denmark, ICT familiar girls and boys performed better than their 'ICT-unfamiliar' same gender counterparts on both the CBAS and P&P test, although the size of the difference between familiar and unfamiliar students was stronger for boys than for girls. In Korea, ICT familiarity is also associated with higher scores on both CBAS and PISA for both genders, however this effect is stronger on the CBAS test than on the P&P test and as in Denmark is stronger for boys.

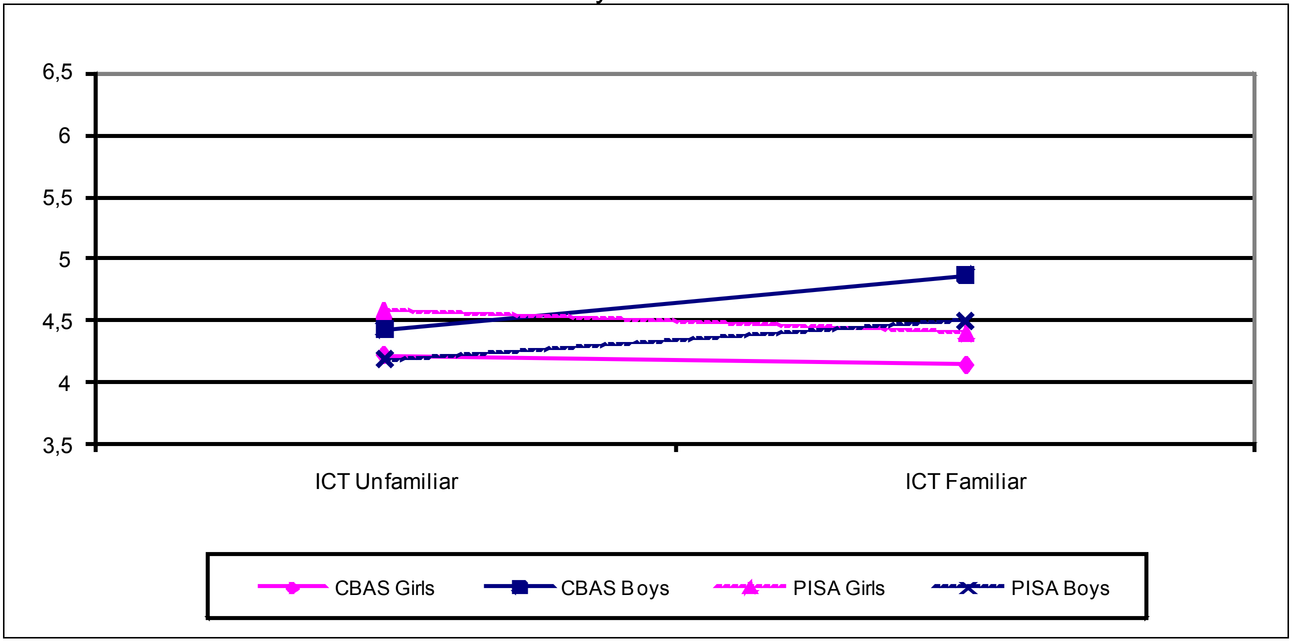


Figure 9: Impact of ICT familiarity for boys & girls on CBAS and P&P scores in Iceland

Figure 9 reveals that in Iceland the same pattern is present for boys, but the reverse pattern is observed for girls: high ICT familiarity is associated with poorer performance for girls on both the CBAS and the P&P tests. On the paper-and-pencil test, ICT unfamiliar girls outperformed their ICT unfamiliar male counterparts. Nevertheless, the reverse was true for the ICT familiar students where boys outperformed girls on the paper-and-pencil test of science and on CBAS.

The Icelandic girls are the only group out of the three countries to display a negative correlation between ICT familiarity and achievement. This may reflect the types of activities that Icelandic girls are performing on computers if these activities are not educational and time spent on the computer replaces other educational activities such as homework or out-of-school lessons. This pattern of results for Icelandic girls requires further investigation in the future to identify what sorts of girls are ICT unfamiliar and why their performance on both the computer-based and the paper-and-pencil test is disadvantaged.

Table 7 shows the size of the advantage for ICT familiar girls and boys in comparison to ICT unfamiliar students and whether these differences were significant or not. The advantage for ICT familiar boys over ICT unfamiliar boys is almost ¼ of a standard deviation, whereas for girls the only significant advantage is for girls in Korea and here the advantage is smaller.

The Icelandic girls stand out here once again where we see that they are the only group for whom there is no trend towards a performance advantage for ICT familiar students. (Although the advantage for the girls in Denmark is also not significant due to the large standard error, a definite trend in this direction is present).

	Advantage for ICT familiar girls	*SE*	*Advantage for ICT familiar boys*	*SE*
Denmark	0.21	0.21	**0.46**	**0.19**
Iceland	0.07	0.10	**0.38**	**0.16**
Korea	**0.29**	**0.13**	**0.43**	**0.16**

Table 7. Effects of ICT familiarity on performance for boys and girls across countries.
*Significant differences are displayed in bold (p<0,05)

Table 8 presents this relationship in another way, displaying the correlations for familiarity and achievement across countries which are stronger for boys than for girls (although on the whole they are relatively weak).

	P&P		*CBAS*	
	Girls	*Boys*	*Girls*	*Boys*
Denmark	0.03	0.09	0.06	0.11
Iceland	-0.04	0.07	-0.02	0.10
Korea	0.05	0.07	0.07	0.10

Table 8. Correlations between ICT familiarity science scores on the CBAS and P&P tests.

Motivation, enjoyment and effort

The CBAS questionnaire was administered so that the relationship between achievement and test engagement factors (enjoyment, motivation and effort) could be investigated. This section examines whether these differences (if any) can explain variations in performance between tests.

The pattern in Figure 10 indicates a clearer trend in Iceland with boys more motivated than girls on the CBAS test. Girls are more likely than boys to strongly disagree or disagree to do the computer-based test “just for fun” whereas boys are more likely than girls to agree or strongly agree. The Fisher’s exact test reveals however that these differences are not significant (FET =5, p>0.05). We note that the patterns of motivation are relatively similar for the paper-and-pencil test of motivation with most Icelandic students disagreeing or strongly disagreeing to do the test “just for fun”. No gender differences are however apparent in motivation for this test (FET =3, p>0.05).

Icelandic students, both boys and girls, are the ‘least motivated’ out of students from all three countries, with the most common response being that they strongly disagree to do another test (regardless of modality) “just for fun”.

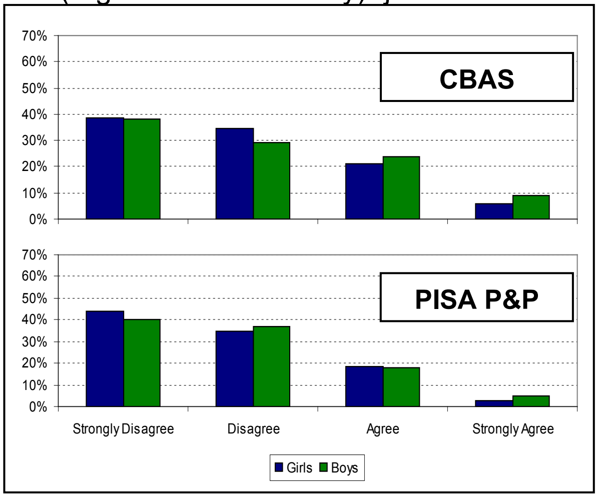


Figure10. Icelandic students’ endorsement of the statement “I would do another computer-based test for fun” (top) and endorsement of “I would do another paper-and-pencil test for fun” (bottom)

In comparing the three countries we noted that Icelandic students overall show less enjoyment of the CBAS and the P&P test compared to students in Denmark and Korea, indicating a specifically cultural pattern of low enjoyment reported by students.

Figure 11 and Figure 12 show that, overall, there is no association between achievement and test motivation or test enjoyment. The only pattern appearing in Denmark seems to be that as boys’ motivation increases, their achievement decreases. In Iceland, again the girls’ response pattern differentiates from all other groups, as they are the only group to show a real association between higher motivation and enjoyment and higher achievement. The boys in Iceland also slightly support this pattern but achievement drops off in the highest category of motivation. In Korea, greater enjoyment is to some extent associated with higher achievement for boys and girls.

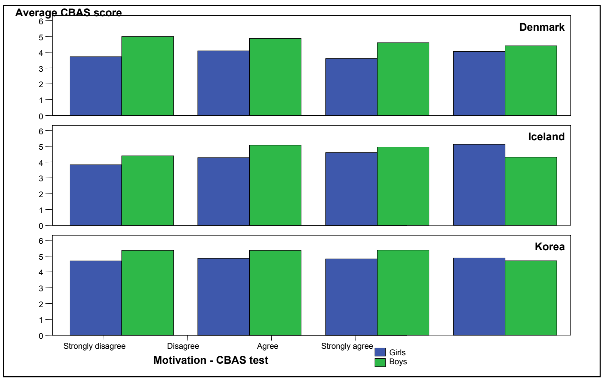


Figure 11: Motivation for CBAS and science achievement

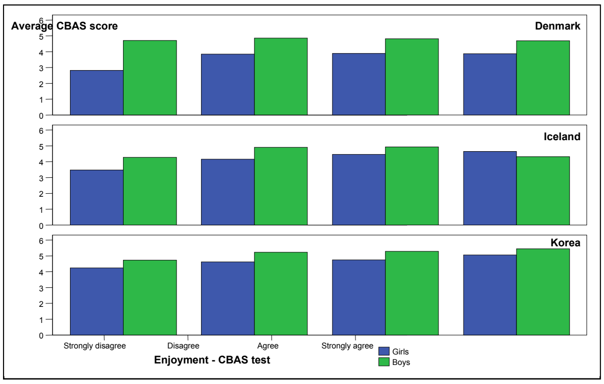


Figure 12: Enjoyment of CBAS test and science achievement

Figure 13 shows the relationship between effort reported on the CBAS Effort Thermometer and CBAS science achievement for boys and girls across the three countries. Only effort thermometer scores with at least five percent of overall responses are displayed (from 5/10 upwards). The figures show only a slight tendency towards higher achievement as reported effort increases across all three countries.

In contrast, as Figure 14: shows, the relationship between PISA P&P reported effort and PISA P&P science achievement is clearly shown as a positive relationship; achievement increases with reported effort for both boys and girls across all three countries.

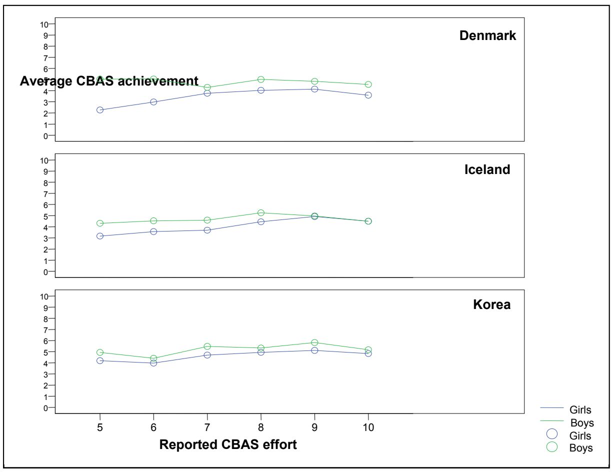


Figure 13: CBAS reported effort and science achievement across countries

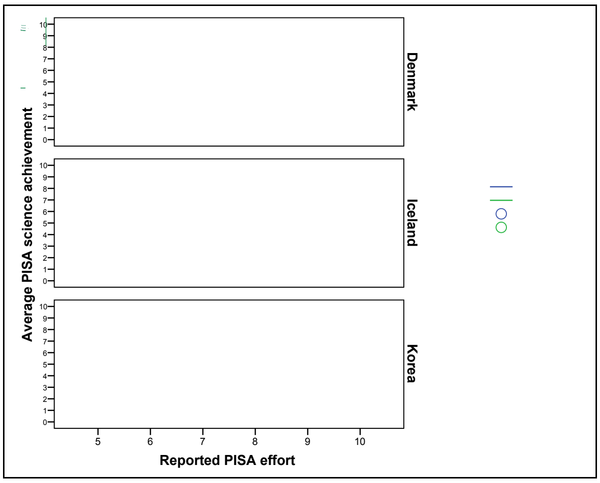


Figure 14: PISA reported effort and PISA science achievement across countries

The correlation data in Table 8 confirm these trends, showing that for the P&P test, if a student reported that they had put a lot of effort into the test, this was associated with higher performance across all countries and for both boys and girls. For the CBAS test this relationship was much weaker, particularly for the boys.

	P&P effort And achievement			*CBAS effort and achievement*			*CBAS report effort and P&P reported effort*		
	Girls	*Boys*	*Total*	*Girls*	*Boys*	*Total*	*Girls*	*Boys*	*Total*
Denmark	0.40	0.28	0.32	0.14	-0.05	0.01	0.41	0.48	0.46
Iceland	0.42	0.42	0.42	0.31	0.17	0.23	0.37	0.31	0.34
Korea	0.26	0.25	0.25	0.12	0.11	0.12	0.60	0.52	0.55

Table 8: Correlations between reported effort and achievement

Interactivity

One explanation of the gender difference in performance proposed is that boys outperform girls on the computer-based items because they are more competent in the types of ICT tasks required of them to complete the items due to their greater ICT familiarity. To investigate this proposal the CBAS items were categorised in terms of their degree of interactivity – for example, whether the item required specific ICT skills such as dragging and dropping or whether it was a relatively simple item involving watching a video and clicking in a response box.

To investigate this, percentage correct was compared for high interactivity items and for low interactivity items. As Figure 15 shows, overall, across both genders and in all three countries, high interactivity items were more difficult than low interactivity items. However, there was no evidence to suggest that the low interactivity items were relatively easier for the girls than for the boys. This is in contrast to the PISA 2003 ICT report (OECD, 2005b) that showed that the more advanced the ICT tasks became, the wider the gender gap. The absence of a gender gap here indicates that the types of ICT skills necessary to answer these questions are relatively low level and well within the grasp of most 15 year old students, both girls and boys.

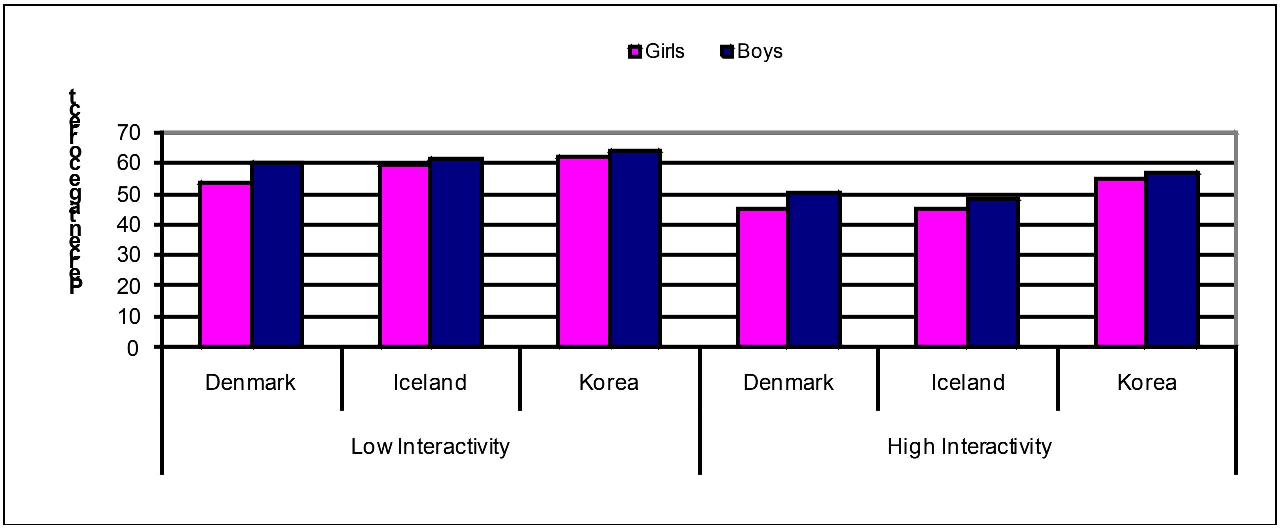


Figure 15: Percentage correct on High and Low interactivity CBAS items

For both genders and in all three countries, high interactivity items were more difficult than low interactivity items. The lower percentage correct overall for high interactivity items may in reality be an artefact of the item type in that the more complex items, e.g. items involving dragging dropping or trialling experiments, required more complex, often two part, answers, calling upon higher reasoning which students had a higher chance of getting wrong.

Reading load

One of the goals of CBAS was to reduce the reading load of the questions, but at the same time retain the science content. It was found that the correlation between the CBAS science and PISA reading literacy, at 0.73, was lower than the correlation between PISA science and PISA reading literacy (0.83), so by this measure the goal of reducing the effect of reading ability was successful. The following analyses investigate the differences in performance on science items varying in degree of reading load for both the CBAS and P&P tests.

All CBAS items were classified as high, medium or low reading load according to the number of words in the item stimulus and question. Percentage correct was calculated for all participating students on the High and Low reading load items. Overall, the higher reading load items were more difficult than the lower reading load items, both for boys and girls across all three countries in both test modalities.

A marked difference in the size of the gender difference between percentages correct on the high and low reading load items was expected. That is, a reduction in boys' advantage on computer items over girls was expected when the items were of a higher reading load, because based on the general PISA trend girls have shown higher competency in Reading literacy (OECD, 2007).

As shown in Figure 16, boys outperform girls on the computer-based items regardless of reading load. This advantage is greater in all three countries for items of low reading load although the size of the advantage on the low reading load items over the high reading load items is relatively small – from under 1% change in Denmark to 3% change in Korea.

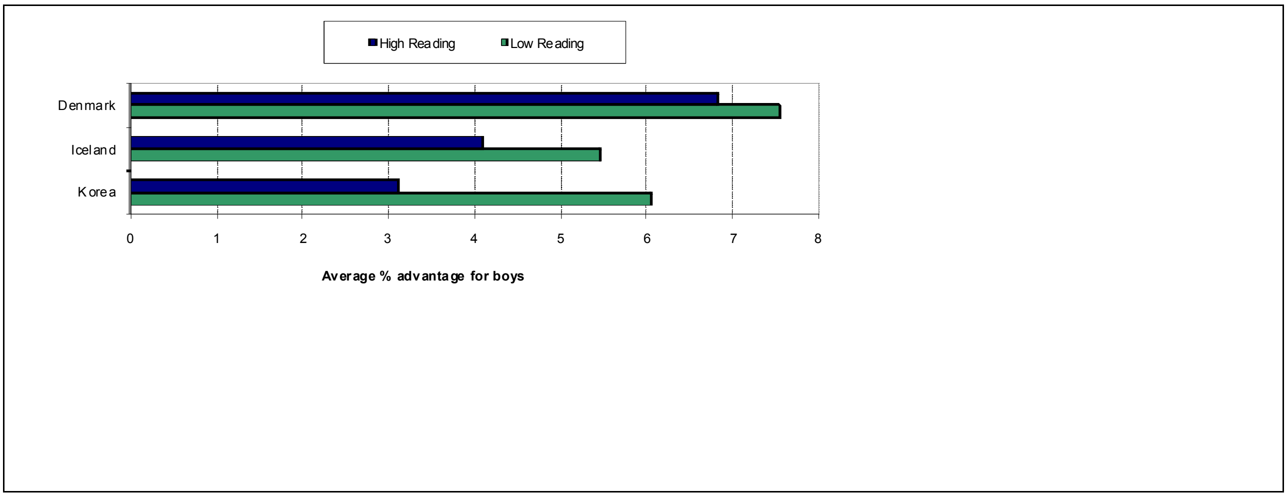


Figure 16: Average percentage difference in achievement between boys and girls on High and Low reading load CBAS items

In Figure 17 the items have been split into three comparison groups: over 200 words, between 100 and 200 words and under 100 words. Note that there are no CBAS items that have over 200 words.

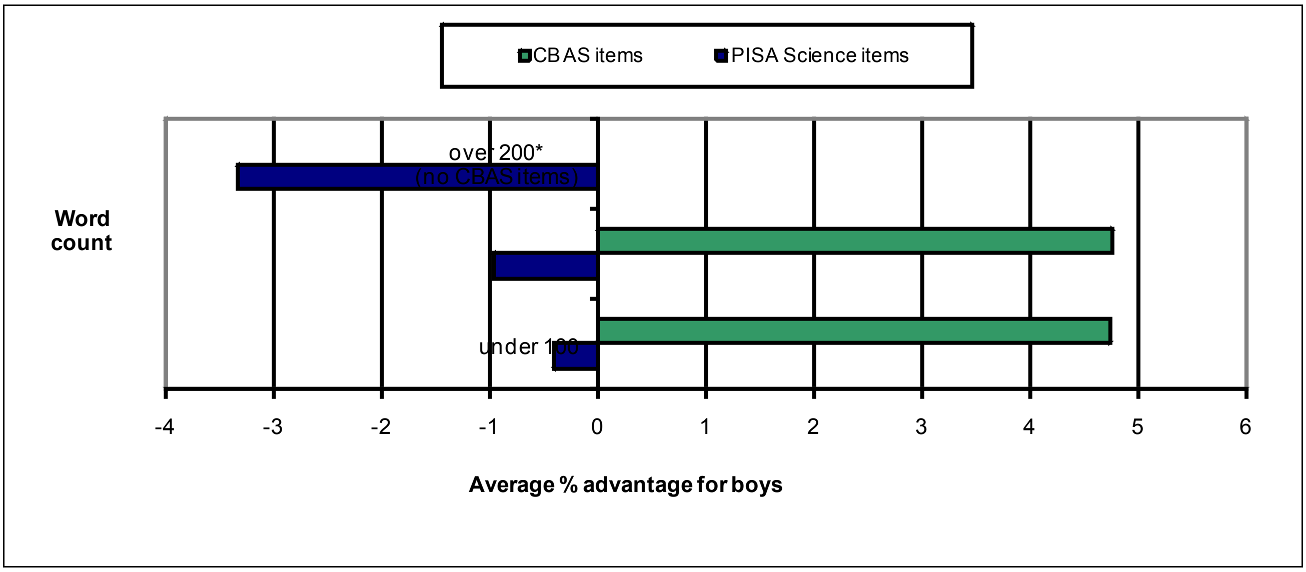


Figure 17: Average percentage difference in achievement between boys and girls according to reading load across test modalities in Iceland

The results in Figure 17 show that, in Iceland, boys outperform girls on the higher and lower reading load CBAS items. Girls outperform the boys on all paper-and-pencil items, but by a far greater degree when the items are long. In fact, when the paper-and-pencil items are similar in length to the CBAS items, the gender advantage for girls is reduced to less than 1% difference. This is consistent with the results from the PISA 2006 assessment (OECD, 2007), considering the overall high reading load of the paper-and-pencil items, where Icelandic girls outperformed their male counterparts by approximately half a standard deviation. These results indicate that boys may be disadvantaged by the length of the paper-and-pencil science items, but they cannot explain fully the advantage for boys on the computer-based items.

Item difficulty

The analyses in this section investigate whether easy CBAS items were comparatively easier for boys or girls. To do this, the PISA item parameters were used for the CBAS items and all CBAS items were classified into three groups according to their item difficulty score: High, Medium and Low. Percentage correct for boys and for girls was calculated for the low and high groups and the average difference between the boys' percentage correct and the girls' percentage was calculated and is displayed in Figure 18:

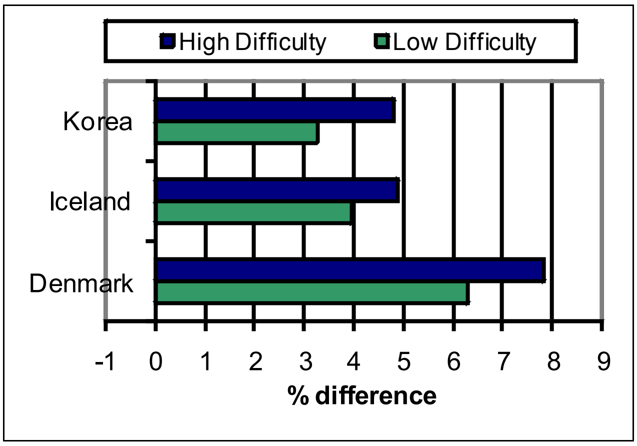


Figure 18: Percentage correct advantage for boys on high and low difficulty CBAS items.

Figure 18 shows that there is a clear advantage in percentage correct for boys in all three countries regardless of the difficulty rating of the CBAS item. Boys' performance advantage is greater for the high difficulty items (they get 5.8% more correct on these items than girls) than on the low difficulty items (where they get on average 4.5% more correct than girls).

While this pattern is notable, it is not, however, unique to the computer-based assessment and a similar pattern is observed in the Icelandic PISA P&P science achievement results. Icelandic boys do comparatively better on the more difficult paper-based science items. Whereas girls outperform boys on the low difficulty items (and overall), there are no performance differences on the high difficulty items. These patterns are shown in Figure 19.

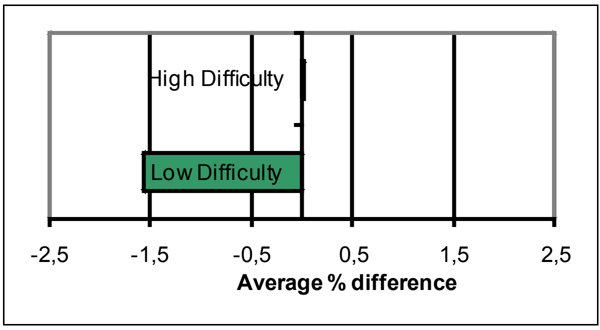


Figure 19. Gender difference in performance in Iceland on harder and easier PISA P&P items

Discussion

By far the most clear cut and most interesting finding from this analysis is the finding that, whereas overall country-by-country performance levels did not change between tests, boys in all three countries outperformed the girls when the test was presented via computer. This gap between the boys' performance and the girls' performance occurred regardless of the patterns of achievement across gender on the PISA paper-and-pencil test of science literacy (recall that in the paper-and-pencil test boys outperformed girls in Denmark, girls outperformed boys in Iceland and there were no gender differences in Korea).

So, Icelandic boys really are better on computerized tests than conventional ones. Then the question becomes why.

The increase in boys' performance may at least partially be explained by the lower reading load. When performance on the paper-and-pencil items that were similar in length to the CBAS items was compared with performance on CBAS there is a substantial increase in performance from the boys and the gender advantage for girls is completely removed.

The questionnaire results appeared to shed light on why the boys perform better on the computer test. In particular,

- boys have more experience with computer-based games, Internet, games-type software that would be similar to the flash animations and video footage used for the CBAS items,
- boys are more motivated on the CBAS test and they enjoy it more,
- boys use computers outside the home more than girls which may contribute to greater confidence in skills transference and greater familiarity with different keyboards, screens and software.

However, despite the intuitive relationship between higher motivation, greater experience with and confidence for ICT tasks and achievement on the computer-based test, statistical analysis of the correlations between achievement and all of these questionnaire factors did not reveal any significant associations between ICT use factors and achievement. Consequently, we must consider other factors that may influence performance as gender differences in performance cannot easily be linked to motivation, computer item interactivity, enjoyment or familiarity with computers.

With regards to interactivity of computer test items gender difference in performance does not clearly increase or decrease according to the interactivity of the items. Boys clearly outperform girls on both high and low interactivity items. However, it should be noted that the 'interactivity' of the CBAS items was relatively low overall as these items were designed to be accessible for even the most ICT unfamiliar students to successfully complete. Also, at the beginning of the CBAS test there was a 10 minute practice session where these response options were demonstrated and practiced. A study such as the PISA 2009 Electronic Reading Assessment with highly interactive items simulating an on-line searching environment will provide researchers with a greater range of item interactivity to examine in more detail potential impacts of interactivity on performance.

Items which show an advantage for boys cannot easily be classified as easier due to low reading load, nor due to higher interactivity. Further, it does not appear to be the domains assessed that affect whether girls or boys will do better on the item, nor the medium of presentation (animation, video footage, still image, etc).

So, *why* are Icelandic boys better on computerized tests than conventional ones? The computer-based items were easier than the paper-based items as the percent correct is much higher overall for CBAS than for P&P science for all countries. The increase in boys' performance in CBAS may partially be explained by lower reading load and by boys' greater test fatigue on low difficulty items. It is possible that the difficulty of the P&P science items fatigues the boys and encourages them to 'give up' more so than girls. This explanation is supported by the finding that gender difference favouring girls in Iceland is removed in performance on paper based items of low reading load (under 100 words). Boys may be disadvantaged by the length of the paper-and-pencil science items.

Reading load cannot explain fully the advantage for boys on CBAS items. Can the rest be explained by a gender bias in the test items themselves?

Was this 'a test for boys'?

Upon closer investigation of the types of items presented in the computer test, it appears that there may be a bias in the gender-typing of the items with a strong content bias towards boys in the video footages used. For example, there are 9 videos over 5 units showing boys performing certain activities, (riding bikes, throwing litter in the bin etc), where the boys are specifically named in the text and sighted in the video footage. There are a further two items in one unit where a boy is named and illustrated as the principal actor in the animated scene. On the girls' side, there is are no items showing girls performing activities by video and only one item that refers to a girl by name. The lack of girls in the items may lead to a lower level of engagement with the test for the girls and a consequently lower level of performance.

Cautionary notes on comparisons of test modality effects

While the overall achievement results were very clear-cut and the gender difference in favour of boys very obvious in each country, finding explanations for the achievement results in the responses to the questionnaire was more difficult due to high levels of variations between countries. The small number of countries involved in this study should be kept in mind when interpreting these results in a wider context and in order to further clarify patterns of changes in performance when testing is presented via computer, further cross-national research will be necessary. For a more balanced test design, valuable insight would be provided in the future by conducting a similar experiment using a third reference group where a group of matched students are given the same paper-and-pencil items via computer.

When analysing modality effects specifically any changes to methods of assessment should be made with caution and preferably after an initial analysis comparing achievement on a paper-and-pencil test with achievement on a computer-based test of the same paper-and-pencil items and achievement on computer-based items in the same domain. In general, changing the test modality to a computer-based presentation platform should not affect performance at the country level; however the current results indicate that it will negatively impact the performance of girls in comparison to the boys. When presenting tests via computer, students may report higher levels of enjoyment, effort and may prefer the computer-based test to a paper-based test but this preference does not mean that achievement will be higher. These domains should be investigated further by national testing institutes wishing to adapt their testing systems, and in particular for Iceland, the reversal of the pattern of achievement by gender and the strange relationship between ICT familiarity and achievement for the Icelandic girls requires detailed future enquiry.

Endnotes

1. See, for example, *New York Times* (January 24, 2005) and *Time Magazine* (March 7, 2005).

Note on this publication:
The article is based on findings from a comprehensive report by the Educational Testing Institude on CBAS for Iceland, Denmark and Korea, scheduled for publication by the OECD in 2009.

References

Harrison, C., Comber, C., Fisher, T., Haw, K., Lewin, C., Lunzer, E., McFarlane, A., Mavers, Di., Scrimshaw, P., Somekh, B., Watling, R. (2004), ImpaCT2: The impact of information and communications technology on pupil learning and attainment, DfES, UK.

Jóhannesson, I.Á. (2004) Karlmennska og jafnréttisuppeldi. [Masculinity and gender equity pedagogy.], Reykjavík: Rannsóknastofa í kvenna- og kynjafræðum [Research Institute in gender and women studies at the University of Iceland]. (In Icelandic)

Kristjánsson, Á.L., Baldursdóttir, S.B., Sigfúsdóttir, I.D. & Sigfússon, J. (2005) Ungt fólk 2004. Menntun, menning, tómstundir, íþróttaiðkun og framtíðarsýn íslenskra ungmenna. [Youth 2004. Education, culture, recreation, sports and youth's visions for the future.], Reykjavík: Rannsóknir og greining. (In Icelandic)

Magnúsdóttir, B.R. (2006). Námshegðun leiðtoga í unglingabekkjum í ljósi rannsókna og kenninga um menningarauðmagn [Educational behaviour of leaders in lower secondary classes and research and theories of cultural capital.], Tímarit um menntarannsóknir, 3, pp. 42–59. (In Icelandic)

OECD (2005a). PISA 2003 Data Analysis Manual: SPSS Users, OECD, Paris.

OECD (2005b), Are Students ready for a Technology-Rich World? What PISA studies tell us, OECD, Paris.

OECD (2007). PISA 2006: Science competencies for tomorrow's world, Vol. 1: Analysis. OECD, Paris.

Ólafsson, R.F., Halldórsson, A.M. & Björnsson, J.K. (2006). Gender and the Urban-rural Differences in Mathematics and Reading: An Overview of PISA 2003 Results in Iceland, in J. Mejding, J. & A. Roe (Eds) Northern Lights on PISA 2003. Copenhagen: Nordic Council of Ministers.

Ólafsson, R.F., Halldórsson, A.M., Skúlason, S. & Björnsson, J.K. (2007) Kynjamunur í PISA og samræmdum prófum 10. bekkjar. [Gender difference in PISA and the National Standard Tests for 10th grade.] Reykjavík: Námsmatsstofnun [Educational Testing Institute]. (In Icelandic)

Ravitz, J., Mergendoller, J. & Rush, W. (2002, April). Cautionary tales about correlations between student computer use and academic achievement. Paper presented at annual meeting of the American Educational Research Association. New Orleans, LA.

Sigfúsdóttir, I.D. (2004) Kynjamunur í skólastarfi [Gender difference in schools.], Uppeldi 3(17), pp. 34–36. (In Icelandic)

Valentine, G., Marsh, J., Pattie, C. & BMRB (2005), Children and young people's home use of ICT for educational purposes: The impact on attainment. Department for education and skills research report RR672, University of Leeds, UK.

The authors:

Almar M. Halldórsson (almar@namsmat.is)
Pippa McKelvie (pippamckelvie@hotmail.com)
Júlíus K. Björnsson (julkb@namsmat.is)
Educational Testing Institute
Borgartún 7A
105 Reykjavík, Iceland

Almar Miðvík Halldórsson, Educational Testing Institute, Iceland. He is the National Project Manager for PISA in Iceland, since 2003. He is also working towards a Ph.D. in Education, evaluating the Icelandic education system from an international perspective in the context of comparative approaches like the PISA, PIRLS and TALIS projects. He is a member of the Icelandic Educational Research Association *(IERA)*.

Pippa McKelvie, has a Bachelor (Honours) degree in Psychology and French from Victoria University of Wellington, New Zealand, and a Master of Science degree in Psychology from Melbourne University, Australia. From 2005 to 2007 she worked as a researcher at the Australian Council for Educational Research on the international surveys: TIMSS and PISA.

Júlíus K. Björnsson, director of the Educational Testing Institute in Iceland. The institute oversees and manages the national tests for primary and lower secondary grades, handles international surveys at these levels, such as PISA, PIRLS and TALIS and is involved in standardisation of tests, both educational and psychological, for example WISC and WPPSI. Júlíus is a member of the PISA Governing Board.

CBAS in Korea: Experiences, Results and Challenges

Mee-Kyeong Lee
Korea Institute of Curriculum & Evaluation (KICE), Korea

Abstract
Korea was one of three countries that participated in the Computer-based Assessment in Science (CBAS) of PISA 2006. The difficulties and advantages of the CBAS implementation are described here based on our experiences made. The CBAS item characteristics and the CBAS results for Korea, including gender differences and attitudes toward both computer-based and paper-and-pencil tests are presented. Finally, implications and challenges for further efforts to move from paper-and-pencil tests to computer-based tests will be discussed. A conclusion that can be drawn from the CBAS experience is that the transition from the paper-and-pencil test to the computer-based test should be made very cautiously, despite the potential advantages of the computer-based test.

The Organization for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA) 2006 survey included the computer-based assessment in science (CBAS) as an option that PISA participating countries could choose, whether or not they participated in that program. PISA is an international assessment that measures the literacy of 15-year-old students in reading, mathematics, and science every three years. PISA measures one domain in depth in each cycle: Reading was a major domain in PISA 2000, mathematics in PISA 2003, and science in PISA 2006.

As science became the major domain for PISA 2006, a computer-based assessment was developed to assess scientific literacy and to provide in depth information about the scientific literacy of 15-year old students. CBAS was designed to add value by enhancing coverage of the PISA Scientific Literacy Framework through instruments that featured a reduced reading load, compared to the paper-based instruments, and dynamic and interactive stimulus materials (OECD, 2005). Introducing computer-based assessment in a large-scale assessment such as PISA is challenging and meaningful because it shows the possibility of using computer-based assessment in large-scale testing situations. Computer-based assessment is becoming increasingly important in the educational assessment area because of its potential for providing high-quality educational assessments. Some examples of the advantages of computer-based assessment include the ability to offer interactions with test items, to provide real life contexts by using dynamic visuals and sound, and to enable near real-time score reporting. Educational professionals expect those features to be increasingly used to improve educational assessments.

In PISA 2006, scientific literacy was defined as an individual's scientific knowledge and use of knowledge, understanding of the characteristics of science as a form of human knowledge and inquiry, awareness of how science and technology shape our material, intellectual and cultural environment, and willingness to engage with science-related issues, and with the ideas of science, as a reflective citizen (OECD, 2007). In addition, the PISA 2006 Science Framework required student competencies in identifying scientific issues, explaining phenomena scientifically and using scientific evidence. Those competencies are abilities required in scientific inquiry, the process of science. Scientific inquiry includes making observations; posing questions; examining books and other sources of information to see what is already known; planning investigations; reviewing what is already known in light of experimental evidence; using tools to gather, analyze, and interpret data; proposing answers, explanations, and predictions; and communicating the results (p. 23 NSES, 1996). The paper-and-pencil test is limited in its ability to assess scientific literacy, and especially to assess scientific inquiry, because scientific inquiry abilities are more accurately assessed in the process of solving problems. Computer-based assessment is expected to improve the assessment of scientific inquiry because it can allow students to make observations, manipulate variables, perform examinations, and gather data, tasks that are not possible with the paper-and-pencil test.

CBAS was initially implemented in a field trial by 12 countries with high interests in CBAS among PISA participating countries. However, the main study was implemented by only three countries: Korea, Denmark, and Iceland. The primary reason that only three countries remained for the main study is that the implementation process required to ensure comparability of results was very complex and strict. CBAS required high standards in the areas of test development, instrument translation and uniformity of testing conditions, and the standards were not easy to follow. The uniformity of testing conditions was the most difficult aspect.

Korea has been participating in the PISA surveys since PISA 2000 and also participated in CBAS in PISA 2006. Unfortunately, the CBAS results cannot be compared to the PISA 2006 paper-and-pencil test results because of the differences of the test items. This makes it difficult to discern the effects of a change in test delivery methods on students' science achievement. However, CBAS suggests implications for efforts to introduce computer-based assessment in a larger-scale test. This study examines the Korean results from the CBAS main study and discusses the implications and challenges for further study.

Implementation of Computer-Based Assessment in Science (CBAS)

Korea implemented the computer-based assessment in science (CBAS) main study from June 12-30, 2006. Seventy-nine schools were sub-sampled from the 154 schools that participated in PISA 2006, and 20 students per school from the 79 schools (1532 total) were sub-sampled from the PISA student sample. Out of the 1532 sampled students, 1485 responded, a 96.9% response rate.

Note that the CBAS data used in the study includes all students who participated in the CBAS test session, as well as all PISA participating students from the schools that had at least one CBAS participant (McKelvie, Halldórsson & Björnsson, 2008). As mentioned, 1532 Korean students were sampled, but the analysis used the data for 2650 students because it included students who participated in the PISA 2006 paper-and-pencil test of science, attended a CBAS-participating school but did not participate in the CBAS test.

Instrument

The CBAS instrument includes 45 science items and survey questions on student attitudes toward the test. But two science items were dropped prior to the analysis, resulting in 43 items being used. All science items were either multiple choice or complex multiple choice. Attitude questions were about enjoyment, motivation about both the paper-and-pencil and the computer-based tests, and test preference. All items were developed in English, translated into Korean and verified with the OECD consortium.

Test administration

CBAS sessions were administered within three weeks after the PISA paper-and-pencil test were implemented. OECD stressed providing participating students with the same environment for CBAS to minimize other effects and influences. Providing the same kinds of laptops was the most important requirement in order to maintain the same test conditions from school to school.

For CBAS administration, we employed 10 Test Administrators and rented 55 laptops and six cars. All 10 Test Administrators had backgrounds in education and information technology. The Korean National Center for PISA provided two training sessions for the test administrators, divided them into five groups and made them travel all around the country with 11 laptops and other materials, such as headsets. One Test Administrator group visited 25 to 30 schools for about three weeks, an average of two schools per day.

The CBAS testing period was for one hour, but the total time for implementation per school was about three hours, which included time for breaks and setting and finishing up, because we usually administered two sessions in each school.

1st hour (1st group: 10 students)	Introduction to CBAS	10 min
	CBAS test	60 min
Break		20 min
2nd hour (2nd group: 10 students)	Introduction to CBAS	10 min
	CBAS Test	60 min
Finishing up		20 min

Table 1: Time table for CBAS

Difficulties and advantages in implementing CBAS

There were some difficulties and advantages in implementing CBAS. The difficulties we experienced included high cost, problems in recruiting quality Test Administrators, and complaints from the participating schools and students. The cost was high because we had to rent laptops and cars. In addition, we had to employ the Test Administrators for one month because the Korean National Center for PISA did not have enough human resources. The Test Administrators had to visit all the schools sampled in order to administer CBAS. Recruiting the Test Administrators was not easy because of the specific requirements for and limitations of the position: We required the Test Administrators to have backgrounds in education and computers, as well as a driver's license. There were not many quality people who could be employed for just one month. Another difficulty was that we received many complaints from the participating schools and students who did not want to participate in the PISA tests twice in such a short time period. We had to spend time persuading them to participate in CBAS.

The advantages we found from the CBAS implementation are as follows: First, we could assess the science competencies that could not be assessed in the paper-and-pencil test. Second, students were more motivated to take CBAS than the PISA paper-and-pencil test. The staffs in the Korean National Center randomly visited both the CBAS and the paper-and-pencil sessions and could see that more students were concentrated in the CBAS sessions. Third, we did not need to print test booklets and code student answers. Finally, CBAS resulted in increased interest in using the computer-based test in a larger test situation in Korea. CBAS certainly drew some interest from policy makers and researchers in Korea.

Characteristics of the CBAS Items

The main characteristics of the CBAS items are adding value to science assessment, reducing the reading load of the test, and minimizing the Information and Communication Technology (ICT) skills required for CBAS (OECD, 2005).

Adding value to science assessments by utilizing the strengths of computers was one of the purposes of CBAS. By allowing students to interact with the computer, CBAS added value to science assessments. For example, students collected data by observation and controlled variables to answer some of the questions in CBAS sessions. Additionally, some CBAS items provided real-life contexts that student encounter in their everyday lives through such stimuli as simulations and videos. Those features of CBAS made it more authentic than the paper-and-pencil test and allowed us to assess aspects of science unavailable through the paper-and-pencil tests.

The purpose of reducing the reading load was to minimize the influence of reading ability in science assessment. The amount of text in the CBAS items is smaller compared to the PISA paper-and-pencil test in science. The stimuli in most of the CBAS items were provided by using videos or animations instead of text.

The reason for minimizing the required ICT skills was also to reduce influences outside of science abilities. To achieve that, CBAS simplified the types of interaction: No keyboard responses, no scrolling, no hyperlinks. Only the skills to click radio buttons, click navigation, and drag and drop were required. A practice session was also included to allow students to become familiar with CBAS.

The CBAS items were developed based on the PISA 2006 Science Framework. However, it is hard to say that CBAS assessed the same literacy as the PISA 2006 paper-and-pencil test in science measured, because the items were quite different. The CBAS item distributions are shown in Table 2 and Table 3.

Science competencies	Number of Items (%)
Identifying scientific questions	10 (23.3)
Explaining phenomena scientifically	18 (41.9)
Using scientific evidence	15 (34.9)
Total	43 (100)

Table 2: Distribution of the CBAS items by science competencies

In the PISA 2006 Science Framework, science competencies consisted of identifying scientific questions, explaining phenomena scientifically and using scientific evidence. The category of questions for explaining phenomena scientifically has the highest number of items, with the category for identifying scientific questions having the lowest number.

The CBAS items can also be classified by the type of knowledge being assessed: Knowledge of science and knowledge about science. The percentages for questions about knowledge of science and knowledge about science are similar: 51.2% and 48.8%, respectively. For the items relating to knowledge of science, the physical systems category has the most items, and the Earth & space systems category has the fewest.

Knowledge	Number of Items (%)
Knowledge of science	22 (51.2)
Physical systems	10 (23.3)
Living systems	7 (16.3)
Earth & space systems	5 (11.6)
Knowledge about science	21 (48.8)
Scientific explanation	9 (20.9)
Scientific enquiry	8 (18.6)
Science & technology	4 (9.3)
Total	43 (100)

Table 3: Distribution of the CBAS items by knowledge

The CBAS Results for Korea

The Korean students showed the highest achievement in CBAS among the three participating countries. The CBAS mean scores were 504 in Korea, 463 in Denmark, and 472 in Iceland. The Korean students also showed higher achievement in the PISA 2006 paper-and-pencil test in science. The same students' mean scores in the PISA 2006 paper-and-pencil test were 502 in Korea, 481 in Denmark, and 471 in Iceland (OECD, 2007).

Gender Differences

Table 4 shows the CBAS results for Korea by gender. The mean scores of girls and boys were 489 and 515, respectively. Boys outperformed girls by 26 score points. The gender difference in CBAS is bigger than in the PISA 2006 paper-and-pencil test of science and the direction of the difference is opposite: The same girls outperformed the same boys on the PISA 2006 paper-and-pencil test of science, although the difference is very small: 503 for girls, 502 for boys. The results imply that the boys' performance improved, while the girls' performance declined in CBAS. Two possible reasons for that difference could be considered: First, that boys are more familiar with computers; second, the item distribution between CBAS and the PISA 2006 paper-and-pencil test of science is different: CBAS has more physical science items, while the PISA 2006 paper-and-pencil test has more biology items. There is much research reporting that boys are strong in physical science (Georgousi, Kampourakis, & Tsaparlis, 2001; Lee & Burkam, 1996; Martin et al., 2004; OECD, 2007).

	Girls	Boys	Gender difference (Girls-Boys)
Mean scores	489	515	-26
SD	94	102	

Table 4: The CBAS results for Korea by gender

Gender differences by regions were analyzed in the study because the ICT environment is different from region to region in Korea (Hwang & Yu, 2005). Table 5 shows gender differences in CBAS and the PISA 2006 paper-and-pencil test in science by region. In CBAS, boys outperformed in metropolitan and urban areas, while girls outperformed in rural areas. The gender difference in CBAS was greatest in urban areas and lowest in rural areas, in the opposite direction. The different trends for gender difference in each region could reflect the different ICT environments by regions: Both girls and boys in rural areas would be less familiar with ICT than those in other regions, because the ICT environment is not as strong there as in metropolitan and urban areas in Korea. As a result, computer familiarity would not be different between girls and boys in rural areas and does not affect the boys' achievement in CBAS as much as in other regions.

The gender differences in the PISA 2006 paper-and-pencil test in science support this argument. The trend in gender differences between the PISA 2006 paper-and-pencil test and CBAS shows that gender differences are changed in favor of boys in CBAS in all regions, although the amount of change is either less or more different from region to region. In metropolitan areas, the direction of gender difference changed in CBAS. In urban areas, the gender difference favouring boys grew. In rural areas, the gender difference favouring girls decreased. The amount of change in gender difference between the two test methods was the largest in metropolitan areas and smallest in rural areas. This implies that boys in metropolitan areas have the most advantage in CBAS and boys in rural areas have the least advantage in CBAS.

Region	Mean scores					
	CBAS			PISA 2006 Paper & Pencil test		
	Girls	Boys	Diff.	Girls	Boys	Diff.
Metropolitan	486	509	-22	503	497	59
Urban	492	538	-45	506	520	-13
Rural	488	482	6	490	466	23

Table 5: Gender differences in CBAS and the PISA 2006 paper-and-pencil test in science by region

Enjoyment of the computer-based test and the paper-and-pencil test

The students' enjoyment on the computer-based and the paper-and-pencil tests was surveyed by two questions included in CBAS; 'I enjoyed doing the computer-based test' and 'I enjoyed doing the paper-based test'. The percentage of students who strongly agreed or agreed with the statement 'I enjoyed doing the computer-based test' was 46.4%, while the percentage of students who strongly agreed or agreed with the statement 'I enjoyed doing the paper-based test' was 19.8%. About a half of the students enjoyed the computer-based test, while only about one fifth of the students enjoyed the paper-based test.

The percentage of girls who strongly agreed or agreed with the statements was higher than boys in both questions, but the difference was small: 2.3% for the computer-based test and 1.0% for the paper-based test. This result is different than what was expected. It was expected that more boys than girls would enjoy the computer-based test, because recent research shows that boys are more interested in computers and ICT (Hakkarainen et al., 2000; Horne, 2007; Papastergiou & Solomonidou, 2005).

The percentage of students who did not enjoy the paper-based test was higher than that of students who did not enjoy the computer-based test. Only 8.5% of students did not enjoy the computer-based test, while 35.2% of students did not enjoy the paper-based test. There could be various possible reasons for these results. First, CBAS involved items using visual effects and interactivity with the examinee. Second, students were curious because they were not familiar with the computer-based test. Third, the CBAS items were easier than the paper-based test items. Fourth, the testing period for CBAS was shorter than for the paper-based test: CBAS took one hour and the paper-based test took two hours. Further research is needed to identify the reasons behind the student's greater interest in CBAS.

	I enjoyed doing the computer-based test.			I enjoyed doing the paper-based test.		
	Girls	Boys	Total	Girls	Boys	Total
Agree	30.9	26.5	28.5	16.6	14.9	15.7
Disagree	6.6	5.7	6.1	27.9	24.0	25.8
Strongly disagree	1.3	3.3	2.4	7.5	11.1	9.4
No response or missing	44.4	45.7	45.1	44.4	45.7	45.1

Table 6: The percentages of responses to enjoyment of the computer-based and the paper-based tests

Motivation about the computer-based test and the paper-and-pencil test

The motivation about the computer-based and the paper-and-pencil tests was surveyed by two questions included in CBAS: 'I would do a computer-based science test just for fun' and 'I would do a paper-based science test just for fun'. The percentage of students who strongly agreed or agreed with the statement 'I would do a computer-based science test just for fun' was 17.5%, while the percentage of students who strongly agreed or agreed with the statement 'I would do a paper-based science test just for fun' was 21.1%. This is interesting and inconsistent with the results from the questions measuring enjoyment of CBAS and the paper-based test. Despite more students enjoying CBAS than the paper-based test, more students preferred to do the paper-based science test just for fun. One possible explanation is that students consider taking a test to be a serious matter and might think taking a test is just for studying, not for fun. Therefore, more of them preferred doing the paper-based test because it seems like it is more helpful for studying.

There were also gender differences in the motivation about the computer-based and the paper-based tests. More boys preferred taking the computer-based science test just for fun, while more girls preferred taking the paper-based science test just for fun, although the difference was small. In addition, fewer boys than girls strongly disagreed or disagreed with doing a computer-based science test just for fun. Those results imply that boys prefer doing the computer-based test for fun more than girls do.

	I would do a computer-based science test just for fun (e.g., if it were on the internet or CD-rom, and it gave me the answers).			I would do a paper-based science test just for fun (e.g., if it were in book or magazine, and it gave me the answers).		
	Girls	Boys	Total	Girls	Boys	Total
Strongly Agree	2.7	5.0	4.0	4.5	2.8	3.8
Agree	11.9	14.9	13.5	16.8	17.9	17.3
Disagree	27.7	24.0	25.6	22.4	23.0	22.7
Strongly Disagree	13.3	10.4	11.7	10.6	11.1	11.2
No response or missing	44.4	45.7	45.1	45.7	44.4	45.1

Table 7: The percentages of responses to motivation about the computer-based and the paper and paper tests

Test preference

CBAS also asked students their test preference. More students preferred two hours on computer-based test: 31.4% preferred two hours on computer-based test, and 3.7% preferred two hours on paper-and-pencil test. More boys than girls preferred not only two hours on paper-and-pencil test but also two hours on computer-based test. However, the gender difference was greater in the two hours on computer-based test.

	Test preference		
	Girls	Boys	Total
2 hours on Paper-and-pencil	3.09	4.23	3.7
1 hour each	23.64	16.56	19.7
2 hours on computer-based	28.99	33.25	31.4
No response or missing	44.29	45.64	45.0

Table 8: percentages of responses to test preference

Correlations between enjoyment and achievement, and between motivation and achievement

The correlations between CBAS enjoyment and achievement were weak, and there is a difference between girls and boys: 0.098 for girls and 0.073 for boys. In other words, there is a weak relationship between CBAS enjoyment and achievement for both boys and girls.

The correlations between CBAS motivation and achievement show an interesting pattern, although they are weak. The direction of correlations for boys and girls was opposite: 0.025 for girls and -0.058 for boys. The more the boys are motivated about CBAS, the lower their achievement, while the trend for girls is the opposite, although the relationship is not strong.

CBAS enjoyment and achievement			CBAS motivation and achievement		
Girls	Boys	Total	Girls	Boys	Total
0.098	0.073	0.068	0.025	-0.058	-0.011

Table 9: Correlations between enjoyment and achievement, and between motivation and achievement

Conclusions and Implications

The experience of implementing CBAS was valuable for future attempts to transition to computer-based test in larger scale assessments. Several challenges should be considered for any future attempts to introduce computer-based test into educational assessments.

First, the factors that affect gender differences on the computer-based tests should be identified. Gender differences appeared differently in CBAS as the testing system changed. The pattern of gender differences in CBAS was different from the PISA 2006 paper-and-pencil test. This implies that changing the testing system would affect boys and girls differently. It would not be possible to find the exact factors that caused the different pattern of gender differences in CBAS, because CBAS was different from the PISA 2006 paper-and-pencil test, not only in its testing system, but also in the way the items are presented and the subject knowledge embedded in the items. Further research is required to identify the factors that affect gender differences.

Second, the strengths of the computer-based tests should be maximized when the test items are developed. The computer-based assessment should offer advantages not only for technical perspectives, but also for the assessment of student ability. The strengths of the computer-based tests should be deliberately considered starting with the design stage of the test items. CBAS somewhat achieved the purpose of adding value to the science paper-and-pencil test, in that some of the CBAS test items were able to assess scientific inquiry abilities by providing dynamic and interactive stimuli, which cannot be done with the paper-and-pencil test.

Third, more attention should be paid to other variables, such as motivational influences and ICT skills, that affect the reliability and validity of the computer-based tests. The enjoyment of and motivation about CBAS was different for boys and girls, and although small, there was a relationship between those attitudes and achievement.

Fourth, the administration procedure should be more efficient for a widespread use of the computer-based test in education. The CBAS test administration procedure was complicated to follow and very costly. This difficulty in the test administration made many countries drop their participation in the main study. Simpler and easier procedures, along with providing the same conditions for all examinees, are necessary for more widespread use of the computer-based tests in education.

Finally, efforts to improve the computer-based test should be made to increase the quality of assessments in education. Introducing computer-adaptive tests in a larger scale assessment and solving technical problems, such as different screen size and resolution of different types of computers, should be part of those efforts. The computer-adaptive tests offer more advantages than the computer-based tests, in that the computer-adaptive tests provide individualized tests for each examinee based on their ability level. And there are still many technical issues to be solved related to the computer-based test.

A conclusion that can be drawn from the CBAS experience is that the transition from the paper-and-pencil test to the computer-based test should be made very cautiously, despite the potential advantages of the computer-based test. Continued efforts are needed to maximize the benefits of the computer-based tests in such varied aspects as item development and test administration.

References:

Georgousi, K., Kampourakis, C., & Tsaparlis, G. (2001) Physical-science knowledge and patterns of achievement at the primary-secondary interface, part 2: able and top-achieving students, *Chemistry Education: Research and Practice in Europe*, 2 (3), pp.253-263.

Hakkarainen, K., Ilomäki, L., Lipponen, L., Muukkonen, H., Rahikainen, M., & Tuominen, T. (2000) Students' skills and practices of using ICT: results of a national assessment in Finland, *Computers & Education,* **34** (2), 103–117.

Horne, J. (2007). Gender differences in computerised and conventional educational tests, *Journal of Computer Assisted Learning*, 23(1), 47-55.

Hwang, J. and Yu, J. (2005) *Development of Indicators on Internet Use and Survey the Indicators.* Korea Information Society Development Institute. Seoul.

Lee, V.E., & Burkam, D.T. (1996). Gender differences in middle grade science achievement: Subject domain, ability level, and course emphasis, *Science Education*, 80(6), 613-650.

Martin, M.O., Mullis, I.V.S., González, E.J., & Chrostowski, S.J. (2004). *TIMSS 2003 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Eighth and Fourth Grades*. Chestnut Hill, MA: Boston College.

McKelvie, P., *Halldórsson, A.M., & Björnsson*, J. (2008) PISA CBAS Analysis and Results. Unpublished report. Educational Testing Institute. Iceland.

OECD (2005) *CBAS Preliminary Field Trial Data Analysis,* Meeting paper presented at NPM meeting.

OECD (2007) *PISA 2006 Science Competencies for Tomorrow's World,* OECD: Paris.

Papastergiou, M. & Solomonidou, C. (2005) Gender issues in Internet access and favourite Internet activities among Greek high school pupils inside and outside school, *Computers & Education,* **44**(4), 377–393.

The author:

Mee-Kyeong Lee
Korea Institute of Curriculum & Evaluation (KICE)
25-1, Samchung-Dong, Jongno-Gu
Seoul (110-230), Korea
E-Mail : mklee@kice.re.kr

Dr. Mee-Kyeong Lee is a research fellow at the Korea Institute of Curriculum & Evaluation (KICE). She was the PISA national project manager of Korea from 2004 to 2007. Her research interests and expertise lie in science education with a particular focus on assessment, curriculum, and teaching and learning.

How did Danish students solve the PISA CBAS items?
Right and wrong answers from a gender perspective

Helene Sørensen & Annemarie Møller Andersen
Danish University of Education, Aarhus University, Denmark

Abstract:
The intention by introducing Computer-based Assessment in Science (CBAS) was to reduce the reading load and to retain the science content. CBAS succeeded in doing this but in the CBAS test results showed a significant gender difference in favour of males. This means that the computer-based test is not gender-neutral to the same extent as the paper-and-pencil test. Instead of that the items appear to be more difficult for girls than boys in all three countries which went through CBAS. Within the Danish context we have classified the different items and compared the results with patterns in girls' and boys' answers. Twelve items were chosen for focus group interviews with two groups of students – three girls and three boys. The analysis shows that the students need other competencies than in the paper-and-pencil test and another problem solving strategy. In the Danish context this may be one explanation for the bigger gender difference in CBAS. The items mediate a gendered impression which influences girls more than boys. This together with the social settings around the test requires a high self confidence to score high in the test.

Background

In the international report *Pisa 2006: Science competencies for tomorrow's world (vol. 1)* are listed some arguments for introducing Computer-based Assessment in Science (CBAS) (OECD, 2007). One goal was to reduce the reading load without making the science content less substantial. Another goal was to explore additional competencies in science than what is possible to explore in the Main Study paper-and-pencil test. In a computer-based test it is possible to give the stimulus as a movie instead of a written story. Furthermore, simulations in the CBAS test give possibilities for testing the students' competencies in the area of knowledge of science and knowledge about science as well as their reasoning competencies in science.

ACER (Australian Council for Educational Research) developed software, which was specifically adapted for the PISA CBAS option addressing the PISA-specific needs of translation and student tracking (Turner, 2008).

All the items in the computer-based test were automatically coded and did not required manual coding. The automatic coding mechanism accommodated multiple choice, complex multiple choice, short numeric response, and "drag and drop" response types. Most of the CBAS items contained multimedia elements. CBAS was administered in the PISA Field trial in thirteen countries. The data from this Field Trial was analysed and the final item selection and method of testing was decided. However, only three countries decided to take part in the CBAS Main Study (Denmark, Iceland and Korea).

The data from Main Study was analysed in the report "PISA CBAS analysis and results - Science performance on paper-and-pencil and electronic tests" by Pippa McKelvie, Almar M Halldórsson and Júlíus K. Bjørnsson (McKelvie, Halldórsson, & Bjørnsson, 2008).

The key findings in this report are:

- No differences for countries overall between science achievement on the paper-and-pencil test and science achievement on the computer-based test.
- Boys outperform girls on the computer-based assessment of science in all countries.
- The gender differences in performance cannot easily be linked to motivation, enjoyment or familiarity with computers.

In this paper possible explanations for the gender difference in the Computer-based Assessment in Science pilot study will be put forward and discussed.

Differences between girls´ and boys´ achievement in PISA CBAS

A report on the Norwegian Field trial of both PISA paper-and-pencil and the computer-based test compare students´ achievement in the two tests (Turmo & Lie, 2006a, 2006b).

Table I compares the girls´ and boys´ achievement in the two tests. The results are standardized by applying 10 as the mean and 2 as the standard deviation for each of the tests. Both tests reveal a difference between sexes in favour of boys, but the gap is far larger for the computer-based test.

The difference between the sexes in the paper-and-pencil test is not statistically significant. However, it can be concluded that the girls score significantly lower in the computer-based test than in the paper-and-pencil test. The boys score significantly higher than the girls in the computer-based test.

	Results; PC test	Results; paper test
Girls (N=157)	9.71	10.26
Boys (N=149)	10.47	10.38
Difference in favour of boys	0.76	0.12
Effect size (difference as percentage of the standard deviation)	38	6

Table 1: Gender differences in achievement in the Norwegian CBAS Field Trail study, from Turmo and Lie (2006a p 4)

Turmo and Lie bring forward that the science items in the CBAS field trial are distributed differently to science sub-domains, with nearly half the items in the physics field, compared to the paper-and-pencil test, where half the items were in the biology field (Turmo & Lie, 2006a, 2006b). They argue that this may be the cause for some of the differences between girls' and boys' achievement in the computer-based test.

The results from the main study of PISA CBAS are reported by McKelvie et al. (McKelvie et al., 2008). A summary of the results are:

- *Overall achievement within countries did not change from one test modality to the next (Yet, there was a tendency for Denmark's performance to decrease on the computer-based test).*
- *Korean students outperformed Danish and Icelandic students in the computer-based test just as they did in the paper-and-pencil test.*
- *Boys' performance increased in Iceland and Korea in the computer-based test while girls' performance decreased.*
- *Boys outperformed girls on the computer-based test in all three countries.*
- *Girls outperformed boys on the paper-and-pencil test of science in Iceland whereas there was a gender difference in favour of boys in the paper-and-pencil results for Denmark.*
- *The association between reading literacy and achievement on the science test was less strong for the computer-based items than for the paper-and-pencil items (McKelvie et al., 2008 p 21).*

In the report *Science Competencies for Tomorrow's World*, Vol. 1 (OECD, 2007) it is mentioned that CBAS Main study showed a significant gender difference in favour of males. The three countries, which took part in the CBAS main study, had a difference between sexes in favour of male students (measured in the score points with 500 as mean and with SD as 100):

- Denmark 45
- Iceland 25
- Korea 26

However, McKelvie et al. found that to be able to compare between the students' achievements across the two test modalities, it was necessary to rescale. Therefore, they constructed a scale with a mean value of 5 and a Standard Deviation of 2 (99% of students have scores between 0 and 10) (McKelvie et al., 2008 p 21).

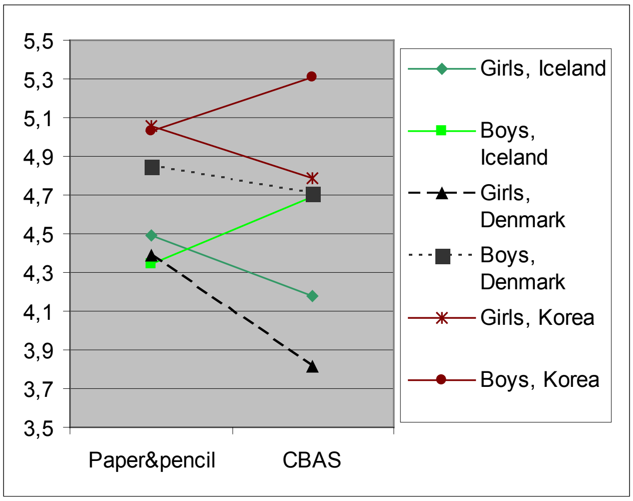


Figure 1: The test scores in Paper-and-pencil and in CBAS test. Rescaled and redrawn from McKelvie et al. (2008 p 25)

Figure 1 shows that boys perform better than girls in the computer-based test in the three participating countries. That indicates that it is not simple to change test modality. Both McKelvie et al and Turmo and Lie have found that because of the lower reading load and a more positive attitude toward the computer-based test, this test modality favours boys:

The two tests produce rather different results, not least viewed from a gender perspective. The PC test clearly favours the boys, and the main explanation seems to be that the reading load is lower, and that the boys are more motivated and active towards the PC-based test. Analysis of the students' active use of multimedia elements shows an interesting gender-related pattern. Also, low-performing male students are active towards the media, while among females this is more a characteristic of the high-performing students (Turmo & Lie, 2006a p 9).

However, McKelvie et al. also mention that "the gender differences in performance cannot easily be linked to motivation, enjoyment or familiarity with computers" ((McKelvie et al., 2008 p 7).

But other explanations may explain the unexpected large difference in achievement between sexes.

Background for the analysis of the CBAS items

Previous studies have demonstrated that assessments in science do mediate differently to girls and boys. When a test item is set in an everyday context the girls tend to stick to the context. For example, they take the decoration of a house into consideration if the task is to plan the electricity circuits (Murphy, 1995; Murphy & Whitelegg, 2006; Sørensen, 1990, 1991; Turmo, 2005).

Girls are more influenced by context, thus familiarity with the content in the item play a greater role for girls than for boys. In the Trends in International Mathematics and Science Study (TIMSS) in 2003 it was found that Norwegian girls in grade 8 scored somewhat higher than boys in biology, while boys scored higher than girls in chemistry, physics and earth science (Grønmo, Bergem, Kjærnsli, Lie, & Turmo, 2004).

Different weighting of the science sub-domains may be expected to affect the size of the gender difference in achievement.

Therefore, to be able to compare the two test modalities, we categorised the sub-domains in CBAS and compared this with the sub-domains in the paper-and-pencil test. PISA CBAS items used in the main study were categorised in sub-domains as shown in figure 2.

Furthermore, to be able to categorise the CBAS items we found it necessary to define additional competences. 9 scientific competencies were defined.

A set of mathematical competences (Niss, 2003a, 2003b) were redefined with the aim of using them in connection to science. These competencies are mentioned in the following list:

1. *Thinking scientifically*
2. *(Posing and) solving scientific problems*
3. *Modelling scientifically*
4. *Reasoning scientifically*
5. *Representing scientifically entities*
6. *Handling scientific symbols and formalisms,*
7. *Communicating in, with, and about science*
8. *Making use of aids and tools (including information technology)*

We added one science related competency:

9. *Observing science situations*

These competencies were used to classify the CBAS items and to identify the items we used for interviews of students.

Two groups of students answered the paper-and-pencil unit "Acid Rain" and were interviewed afterwards (15 students). Two groups of students (three girls and three boys) solved 12 CBAS units and were interviewed after each item about their problem solving strategies.

The gendered influence of the distribution of sub-domains

A gendered bias in the test items

In the main PISA paper-and-pencil test the distribution between sub-domains was intended to be as shown in table II.

Desired distribution of score point for Knowledge	
Knowledge of science	**Per cent of score points**
Physical systems	15-20
Living systems	20-25
Earth and space systems	10-20
Technological systems	5-10
Subtotal	60-65
Knowledge about science	
Scientific enquiry	15-20
Scientific explanation	15-20
Subtotal	35-40
Total	100

Table 2: The intended distribution of items in sub-domains. after fig. 1.8 (OECD, 2006)

In the PISA main study paper-and-pencil test there was a balance close to the intended distribution of assessing various components of the science literacy framework. There was nearly the same amount of score-points allocated to Living systems as to Physical systems and Earth and Space systems together (Kjærnsli, Lie, Olsen, & Roe, 2007).

For the PISA paper-and-pencil test 2006 the different scores for the science domains are shown in table III.

	Science Scale	Physical systems scale		Living systems scale		Earth and space scale	
		Mean score	M - F	Mean score	M - F	Mean score	M -F
Korea	522	530	15	498	6	533	14
Denmark	496	502	29	505	11	487	26
Iceland	491	493	15	481	-5	503	7
OECD	500	500	26	502	4	500	17

Table 3: Student performance on the science scale and mean score and gender differences (M-F) on the three "knowledge of science" scales. (Values that are statistically significant are indicated in bold).(OECD, 2007)

In the two sub-domains *Physical systems and Earth* and *space systems* there is a significant difference between girls' and boys´ achievements in favour of boys, most evident in Denmark.

Compared to this, Living systems was represented by 18 items and *Physical systems* combined with *Earth and space system* were represented by 22 items in the CBAS main study test (see figure 2),

This may explain part of the difference between sexes.

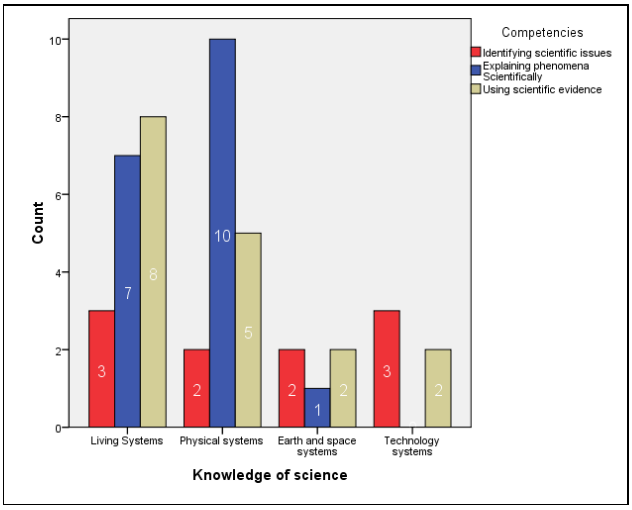


Figure 2: Our categorisation of the PISA CBAS main study items

In CBAS main study nearly one fourth of the items concerned *Explaining phenomena in Physical systems* (see figure 2). Thus, because of the different distribution the CBAS test had a gender bias in favour of boys which may explain the big difference between girls and boys achievement.

For the Danish CBAS data set girls chances for answering right is nearly twice as high as the boys in the sub-domain *Explaining phenomena in Living systems* than in *Explaining phenomena in Physical systems.*

However, this is only one way to evaluate CBAS items in terms of gender. Another explanation could be found when analysing the items. Beside the "boy friendly" distribution of items, it turns out that all persons with an active role in the pictures or video sequences are male, and when persons' names are mentioned all names are male. This is due to the selection of units from the Field trial into the Main study (Adams, 2008), but it gives a gender bias in the test.

The items may also in other ways mediate a gender biased context as the following examples will show.

Mediation of a gendered context

In most of the following discussion we build on the Danish data set. I all three PISA pencil and paper test (2000, 2003 and 2006) Danish boys performed better than Danish girls. These differences are relatively high compared to the other Nordic countries with a similarity in school systems and in culture. We have no simple explanations for this phenomenon. In PISA CBAS main study this difference between girls' and boys' achievement are even higher than in the paper-and-pencil tests.

One hypothesis is that girls and boys in Denmark have different problem solving strategies. Another explanation is, that students are not used to tests in science. We do not in Denmark have tradition of writing in science, and until 2006 there have only been final examination in physical science and this was an oral/practical test.

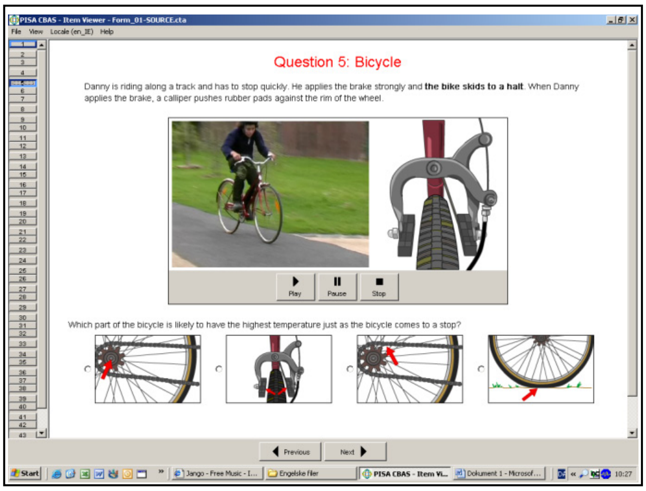
In order to explore the impact of girls' and boys' problem solving strategies, we did interviews involving the PISA paper-and-pencil test item "Acid Rain". We found that the students did not use the stimulus when they answered the test questions. They only tried to answer when they could recall the content from memory.

Especially girls did not try to find an answer when the content was unknown or appeared difficult.

PISA 2006 shows that there are a relative high difference in self-confidence, with Danish boys being more confident (OECD, 2007; Sørensen, 2008; Sørensen & Andersen, 2007), which may explain this behaviour.

Example 1

The video clip shows Danny riding his bicycle and especially how the wheel stops rotating (see figure 3).


PISA CBAS - Item Viewer - Form_01-SOURCE.cta
File View Locale (en_IE) Help
Play
Pause
Stop
Previous
Next
Start

Question 5: Bicycle

Danny is riding along a track and has to stop quickly. He applies the brake strongly and **the bike skids to a halt**. When Danny applies the brake, a calliper pushes rubber pads against the rim of the wheel.

Which part of the bicycle is likely to have the highest temperature just as the bicycle comes to a stop?

○ ○ ○ ○

Figure 3: Bicycle Q1 from PISA CBAS Item viewer (OECD, 2005)

The item "Bicycle question 2" uses the same setup, but this time the wheel still rotates slowly. The text here says: "He applies the brake so that the wheel stops gradually and **the bike does not skid".**
The last sentence in both examples is: "When Danny applies the brake a calliper pushes rubber pads against the rim of the wheel".

Figure 4 shows the Danish students' answers. The answer patterns are the same for the Icelandic and Korean students.

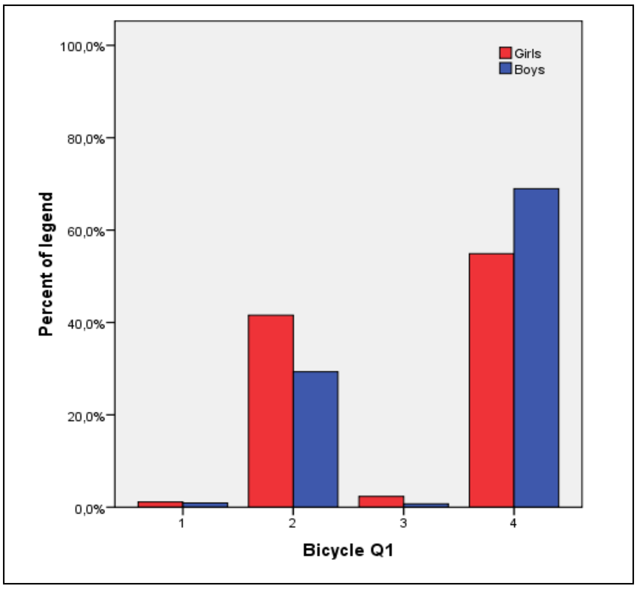


Figure 4a: Danish students' answers to Bicycle Q1

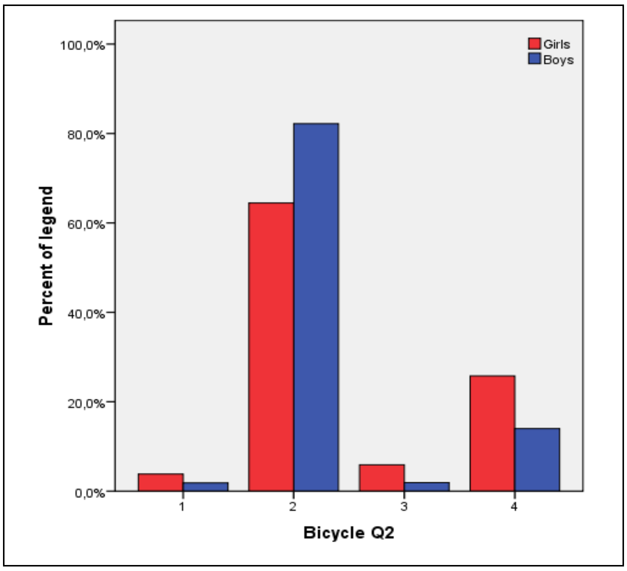


Figure 4b: Danish students' answers to Bicycle Q2

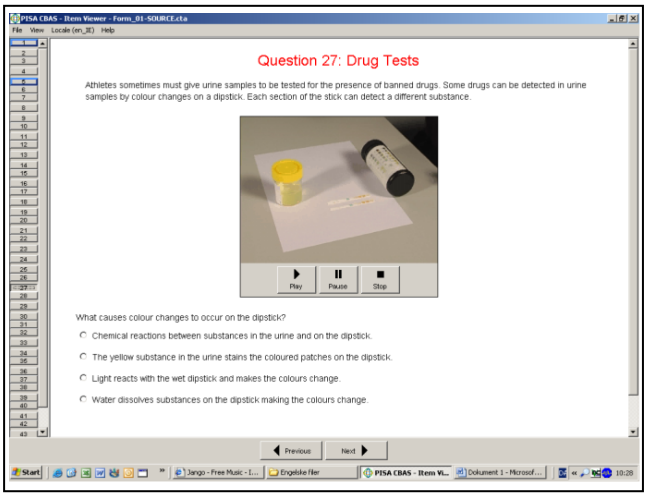
We found that the students needed the following additional scientific competencies, in order to answer the questions:

- *Observing science situations*
- *Thinking scientifically*
- *Solving scientific problem*
- *Modelling scientifically*
- *Reasoning scientifically*

The situations mediate clearly everyday situations and the female students in interview tell that they know what to answer because of that. But to solve the first question, the students need to use scientific thinking and observing. They have to model the situation and reason that all friction takes place between the ground and the tyre, because the wheel does not rotate.

Thus, to be able to give the right answer to these questions the students need knowledge about forces, friction and heat. Those students that stick to the everyday situation and remember that the rubber pads tend to become warm, when you use the brake, will give the wrong answer (answer two instead of four). This was done by 40 % of the Danish girls.

Example 2


PISA CBAS - Item Viewer - Form_01-SOURCE.cta
File View Locale (en_IE) Help
Play
Pause
Stop
Previous
Next
Start
Jango - Free Music - I...
Engelske filer
PISA CBAS - Item Vi...
Dokument 1 - Microsof...
10:28

Question 27: Drug Tests

Athletes sometimes must give urine samples to be tested for the presence of banned drugs. Some drugs can be detected in urine samples by colour changes on a dipstick. Each section of the stick can detect a different substance.

What causes colour changes to occur on the dipstick?

○ Chemical reactions between substances in the urine and on the dipstick.

○ The yellow substance in the urine stains the coloured patches on the dipstick.

○ Light reacts with the wet dipstick and makes the colours change.

○ Water dissolves substances on the dipstick making the colours change.

Figure 5: Drug test Q1 from PISA CBAS Item viewer (OECD, 2005)

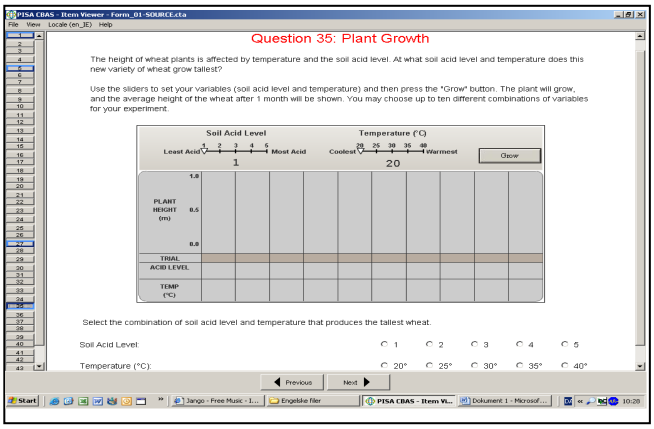
The students also labelled the item shown in figure 5 as a well known situation. This item mediate school science and more than 80 % of the Danish students gave the right answer, both girls and boys (about 60 % in Korea and Iceland). Both girls and boys described in the interviews this situation as well known and sort of everyday. It is part of the Danish school curriculum to measure pH using sticks. Compared to the Bicycle item, this situation clearly communicates that the item is a chemistry problem, thus there is not a confusion about which "rules" are in play.

One could argue that the ability to define the science question from an everyday situation is a part of having obtained scientific literacy. But as mentioned earlier research has shown that girls tended to stick to the everyday context and therefore put value to more variables than boys, which may result in a wrong answer in a science test. This issue has to be addressed explicitly in the classrooms to educate the students. And, in Denmark, learning *about* science is not widespread.

Example 3

The last example consists of two different items – one dealing with *Plant Growth* (figure 6) and the other about *Nuclear Power Plant* (figure 7). Both items deal with simulations.

The competencies involved in both situations are *Thinking scientifically, Modelling scientifically, Reasoning scientifically and making use of aids and tools (including information technology).*



Question 35: Plant Growth

The height of wheat plants is affected by temperature and the soil acid level. At what soil acid level and temperature does this new variety of wheat grow tallest?

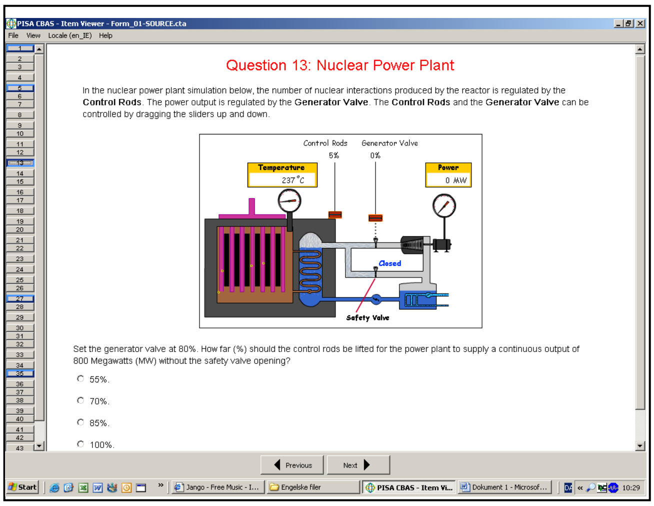
Use the sliders to set your variables (soil acid level and temperature) and then press the "Grow" button. The plant will grow, and the average height of the wheat after 1 month will be shown. You may choose up to ten different combinations of variables for your experiment.

Select the combination of soil acid level and temperature that produces the tallest wheat.

Soil Acid Level: ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Temperature (°C): ○ 20° ○ 25° ○ 30° ○ 35° ○ 40°

Figure 6: Plant Growth Q1 from PISA CBAS Item viewer (OECD, 2005)



Question 13: Nuclear Power Plant

In the nuclear power plant simulation below, the number of nuclear interactions produced by the reactor is regulated by the **Control Rods**. The power output is regulated by the **Generator Valve**. The **Control Rods** and the **Generator Valve** can be controlled by dragging the sliders up and down.

Set the generator valve at 80%. How far (%) should the control rods be lifted for the power plant to supply a continuous output of 800 Megawatts (MW) without the safety valve opening?

○ 55%.

○ 70%.

○ 85%.

○ 100%.

Figure 7: Nuclear Power Q1 from PISA CBAS Item viewer (OECD, 2005)

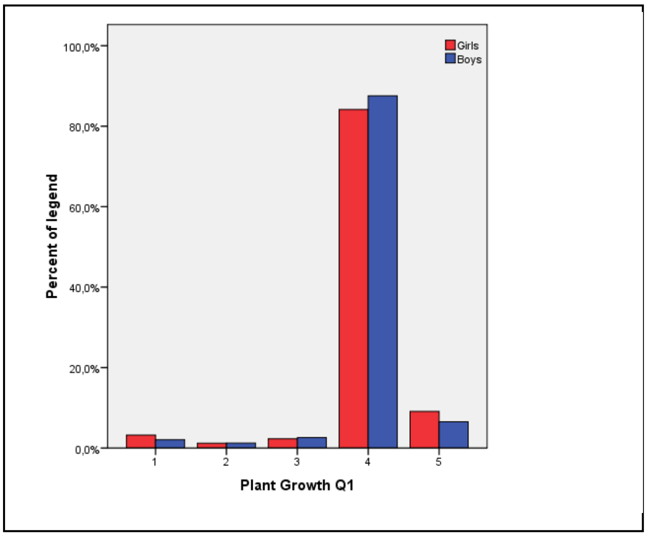


Figure 8a: Danish students' answers to *Plant Growth* Q1

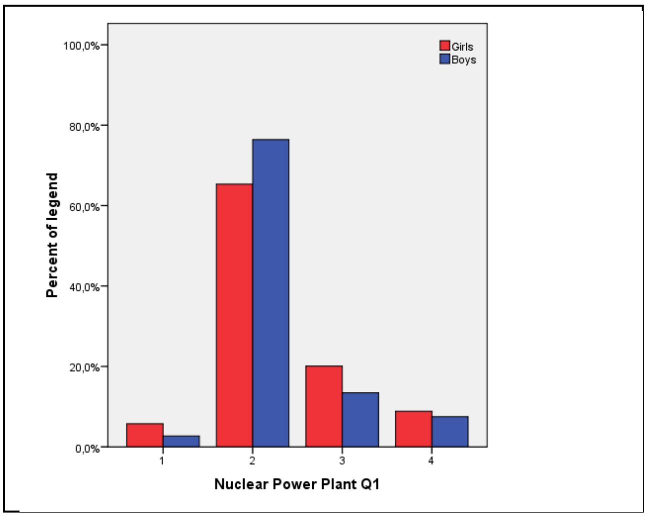


Figure 8b: Danish students' answers to *Nuclear Power Plant* Q1

But the item *Nuclear Power Plant* was perceived as much more difficult by the girls in the interview, because it looked more technical and mediated "difficult physics". The chance for girls' right answer compared with boys was nearly twice as high in *Plant Growth* Q1 as in *Nuclear Power Plant* Q1.

Closing remarks

The use of computer-based testing in science gives possibility for assessing more competencies than it is possible in the traditional paper-and-pencil PISA test. CBAS evaluates competencies more relevant to science education such as:

- Observing
- Reasoning
- Problem-solving
- Modelling
- Manipulating
- Planning experiments

The achievement in science becomes more independent of reading capabilities than in the text heavy paper and pencil.

But to be able to use computer-based testing in science, the issue of favouring the boys in the test has to be addressed in developing new science items.

Overall the PISA CBAS test items mediate a gendered context. Persons and names are male. Several situations appear as "boys playing around".

Research in the field of gender and science has demonstrated that girls are more influenced by context in science assessments. They may not use scientific modelling and reasoning in problem-solving situations but tend to stick to the everyday concepts mediated by the context. The demand for making answering in science terms has to be mere explicit.

Furthermore, items in physics or technology are conceived as more difficult by the girls. The girls find the unfamiliar content more difficult and may than chose not to answer the scientific questions. PISA CBAS has more items in *Physical systems* and in *Technological systems* than the paper-and-pencil PISA 2006. The distribution of the scientific sub-domains has to be considered.

The social setting around the test demands high student confidence both in connection to the test media and concerning the science content. Female students are generally less confident using ICT and the results of PISA 2006 reveal a lather large difference between sexes in self confidence among Danish in favour of boys.

To circumvent this issues the test setup and context has to be considered.

Bibliography

Adams, R. (2008). Personal Communication. Reykavik.

Grønmo, I. S., Bergem, O. K., Kjærnsli, M., Lie, S., & Turmo, A. (2004). *Hva i all verden har skjedd i realfagene? Norske elevers prestasjoner i matematikk og naturfag i TIMSS 2003./2004. Oslo.* Oslo: Department of Teacher Education and School Development, University of Oslo.

Kjærnsli, M., Lie, S., Olsen, R. V., & Roe, A. (2007). *Tid for tunge løft. Norske elevers kompetanse i naturfag, lesing og matematikk i PISA 2006.* Oslo: Universitetsforlaget.

McKelvie, P., Halldórsson, A. M., & Bjørnsson, J. K. (2008). *PISA CBAS analysis and results - Science performance on paper-and-pencil and electronic tests*. Reykjavik: Educational Testing Institute, in press.

Murphy, P. (1995). Problems in practising authentic assessment - the English experience. In A. M. Andersen, K. Schnack & H. Sørensen (Eds.), *Science - Natur/Teknik, Assessment and Learning* (Vol. 22). Copenhagen: Royal Danish School of Educational Studies.

Murphy, P., & Whitelegg, E. (2006). *Girls in the Physics Classroom - a Review of the Research on the Participation of Girls in Physics*: Institute of Physics.

Niss, M. (2003a). *Mathematical competencies and the learning of mathematics: The Danish KOM project.* Paper presented at the 3rd Mediterranean Conference on Mathematical Education, Athens, Greece.

Niss, M. (2003b). Quantitative literacy and mathematical competencies: a Danish perspective. In B. L. Madison & L. A. Steen (Eds.), *Quantitative literacy: why numeracy matters for schools and colleges* (pp. 215-220). Princeton: National Council on Education and the Disciplines.

OECD. (2005). PISA CBAS-Item Viewer-1.4.4. Camberwell, Australia: Australian Council for Educational Research.

OECD. (2006). *Assessing Scientific, Reading and Mathematical Literacy: A framework for PISA 2006*.

OECD. (2007). *PISA 2006: Science Competencies for Tomorrow's World, Vol. 1*. Paris: OECD.

Sørensen, H. (1990). *Fysik- og kemiundervisningen - set i pigeperspektiv.* Danmarks Lærerhøjskole, København.

Sørensen, H. (1991). *Physics and Chemistry in the Danish primary School - seen form the girls´perspective.* Paper presented at the The sixth international GASAT conference, Australia, Melbourne.

Sørensen, H. (2008). A framework for gender inclusive science education. In B. Hodgson (Ed.), *Challenging and changing the Status Quo, Proceedings of the 12th international GASAT conference*. University of Brighton: Gender and Science and Technology Association.

Sørensen, H., & Andersen, A. M. (2007). Elevers holdninger til og interesse for naturfag og naturvidenskab. In N. Egelund (Ed.), *PISA 2007*. København: Danmarks Pædagogiske Universitetsskole.

Turmo, A. (2005). *Gender differences in students' achievement, attitudes, and self-concept in*
science: New evidence from the TIMSS 2003 study in Norway. Paper presented at the ESERA, Barcelona.

Turmo, A., & Lie, S. (2006a). *PISA's Computer-based Assessment of Science (CBAS) – A gender equity perspective*, 2008

Turmo, A., & Lie, S. (2006b). PISA's Computer-based Assessment of Science (CBAS) Gjennomføring og norske resultater våren 2005. *Acta Didacta*(2).

Turner, R. (2008). OECD Programme for International Student Assessment (PISA) 2006 - Computer-based Assessment of Science (International Option), from http://www.acer.edu.au/research_projects/pisa_ict.html

The authors:

Helene Sørensen & Annemarie Møller Andersen
Danish School of Education
Aarhus University
Tuorgvej 124
DK 2400 Copenhagen NV Denmark

Helene Sørensen, Danish School of Education, Aarhus University. She is associate professor and head of the math and science research group. Helene Sorensen is educated as and worked as a teacher in science before doing her Ph.D. concerning gender issues in science education. She has for several years done research in this field. She is member of the Danish PISA group with responsibility for the science part and was member of the evaluation group concerning the Danish National Test.

Annemarie Møller Andersen, Danish School of Education, Aarhus University. She is associate professor (retired) and member of the math and science research group. Annemarie Møller Andersen is educated as and worked as a teacher in biology before doing her Ph.D. in science education. She has done research in different aspects of science education. She was a member of the Danish PISA group with responsibility for the science part until 2008, and she acts as an advisor for the computer-based School-Leaving Examination in biology.

ANNEX: Workshop Programme

Monday, September 29th

09:00 Welcome by the Icelandic Ministry of Education

09:15 *Nathan Thompson & David Weiss* (Assessment Systems Corporation, USA): **Computerized and Adaptive Testing in Educational Assessment**

10:30 **National experiences with computer-based assessment**

- *Brent Bridgeman (ETS, USA):* Experiences from large scale computer-based testing in the USA
- *Jakob Wandall* (Ministry of Education, Denmark): The new Danish National Test

13:00 **The European agenda on international large-scale surveys**

- *Oyvind Bjerkestrand* (European Commission): The European Coherent Framework of Indicators and Benchmarks and implications for computer-based assessment
- *Jostein Ryssevik* (SurveyLang): Large-scale computer-based testing of foreign language skills across Europe – requirements and implementation
- *Ernesto Villalba (JRC, IPSC)*: Computer-based assessment and the measurement of creativity in education

14:30 **Shifting to computer-based assessment**

- *Vesna Busko* (University of Zagreb, Croatia): Shifting from Paper-and-Pencil to Computer-based testing: advantages/disadvantages, consequences for testing outcomes
- *Sam Haldane* (ACER, Australia): Delivery platforms for national and international computer-based surveys: history, issues and current status
- *Thibaud Latour* (Centre de Recherche Public Henri Tudor, Luxemburg): Shifting from Paper-and-Pencil testing : An economic model of Return on Investment (ROI)

16:15 **Experiences with the transition to computer-based assessment**

- *Benő Csapó, Gyongyver Molnar & Krisztina R. Toth* (University of Szeged, Hungary): Comparing paper-and-pencil and online assessment of reasoning skills - a pilot study for introducing TAO in large-scale assessments in Hungary
- *Eli Moe* (Aksis, Unifob, Norway): Introducing large scale computer-based testing of English - experiences and future challenges
- *Heiko Sibberns* (IEA-DPC, Germany): Experiences with moving to computer-based testing: test preparation and field operations in two computer-based assessment tests

17:45 Review of main results

18:30 Reception for all participants at the invitation of the Icelandic Ministry of Education

Tuesday, September 30th

09:00 **CBAS assessment and lessons learned with national representatives from Iceland, Denmark and Korea** (Moderation: Gerben van Lent)

- *Mee-Kyeong Lee*, Korea Institute of Curriculum & Evaluation, Korea
- *Julius Björnsson*, Educational Testing Institute, Iceland
- *Helene Sørensen*, (Aarhus University, Denmark): How do Danish students solve the CBAS items? Looking at right and wrong answers from a gender perspective

10:00 *Benő Csapó* (University of Szeged, Hungary): What is planned in PISA: The ERA component of PISA 2009 and plans for PISA 2012

10:15 Conclusions for future PISA surveys and international computer-based surveys (Moderation: Gerben van Lent)

11:00 **Comparison of traditional and electronic testing: What are the differences?**

- *Gerben van Lent* (ETS Global, Netherlands): Risks and benefits of computer-based testing versus paper-and-pencil
- *Martin Ripley* (Independent consultant, UK): Innovation versus migration
- *Oliver Wilhelm* (IQB, Germany): Traditional and Computerized Ability Measurement: Stressing Equivalence versus Exploiting Opportunities

13:30 **CBAS and comparison of traditional and electronic testing: Gender issues**

- *Romain Martin* (University of Luxemburg, Luxemburg): Gender differences in reading literacy - a consequence of girls' preference for the information presentation formats used in paper and pencil tests?
- *Almar M. Halldorsson* (Educational Testing Institute, Iceland): Are Icelandic boys really better on computerized tests than conventional ones?
- *Martin Ripley* (Consultant, UK): Gender effects in computer-based testing

15.15 **The future of electronic testing: trends, difficulties, obstacles to overcome, technical solutions and research needs** (Moderation: Oliver Wilhelm)

- *Ron Martin* (ACER, Australia): Utilising the full potential of computer delivered surveys in assessing the aims of science teaching
- *Theo Eggen* (CITO, Netherlands): Computer-adaptive testing of basic skills in arithmetic/mathematics for teacher training
- *Patrik Kyllonen* (ETS, USA): New Constructs, Methods, and Directions in Computer-Based Assessment

16:45 *Lara Tilmanis* (Intel Corporation, USA): **Transforming Education Assessment: Teaching and Testing the Skills Needed in the 21st Century / A Call to Action**

17:15 Final comments by the Educational Testing Institute, the JRC and the OECD

Wednesday, October 1st

The new Icelandic National Testing programme

09.00 Introduction – Agenda and purpose of the meeting

09.15 A short introduction to the Icelandic National Tests, - history and recent developments

10.00 Coffee

10.30 The proposal for changing over to electronic testing

11.30 Discussion

12.00 Lunch

13.00 Requirements of the new testing programme: What needs to be done?

14.00 Hardware and software requirements – introduction to the TAO system. What needs to be adapted changed and adapted?

15.00 General discussion – problems and solutions

16.00 Final comments and farewell

Further information about profiles and access to slides is provided at the CRELL web-site of the event:

http://crell.jrc.it/WP/workshoptransition.htm

European Commission

EUR 23679 EN – Joint Research Centre – Institute for the Protection and Security of the Citizen
Title: THE TRANSITION TO COMPUTER-BASED ASSESSMENT New Approaches to Skills Assessment and Implications for Large-scale Testing.
Editors: Friedrich Scheuermann & Julius Björnsson
Luxembourg: Office for Official Publications of the European Communities
2009 – 207 pp.
EUR – Scientific and Technical Research series
ISSN 1018-5593
ISBN 978-92-79-11110-5
DOI 10.2788/60083

Abstract

In September 2008 the Joint Research Centre (JRC, IPSC) of the European Commission, together with the Iceland Educational Testing Institute, carried out an expert workshop on “The Transition to Computer-Based Assessment - Lessons learned from the PISA 2006 Computer-based Assessment of Science (CBAS) and implications for large scale testing”. This report is based on input made and conclusions drawn from the discussions held on computer-based skills assessment in comparative surveys, such as the international PISA survey which is going to be fully computer-based in the near future. Specific emphasis is given to the comparison between paper-pencil tests and computer-based assessment, electronic tests and gender differences, and adaptive vs. linear computer-based assessment. The volume is complemented by articles in areas which have not been covered by workshop presentations. It therefore provides a comprehensive overview of issues and challenges to take into account when moving from traditional testing approaches to computer-based assessments.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

JRC
EUROPEAN COMMISSION

LB-NA 23679 EN C

ISBN 978-92-79-11110-5

9 789279 111105

Publications Office
Publications.eu.int