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	χ2	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

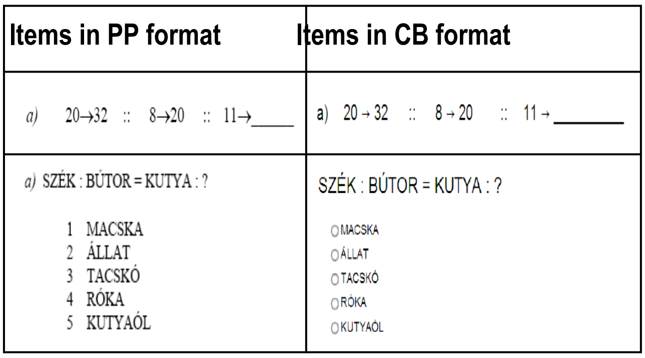
Instruments

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.