Implementing an Online Formative Assessment System: From Paper-Based to Computer-Based Testing

Paper presented at the symposium on
Implementing Large-Scale Technology-Based Assessments in Five Countries,
AERA, 2010, Denver, USA

Benő Csapó, Gyöngyvér Molnár and Krisztina R. Tóth
University of Szeged, Hungary

In the 21st century it is no more doubt that a notable percentage of educational assessment taking place is based on technology, mostly on computers, showing how fast the use of computer-based assessment has increased (see e.g., Csapó, Latour, Bennett, Ainley, & Law, 2009). Most of the research literature on computer-based assessment reflects a predominance of research comparing the results of paper-based and computer-based assessment of the same construct (e.g. see Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008; Horkay, Bennett, Allen, Kaplan, & Yan, 2006; Wang, Jiao, Young, Brooks, & Olson, 2007; Wang, Jiao, Young, Brooks, & Olson, 2008; Csapó, Molnár, & R. Tóth, 2009).

Previous research has indicated that identical paper-based and computer-based tests will not always obtain the same results (Clariana, & Wallance, 2002). Therefore, comparability is important, if one prefers to compare results over time, in which the delivery mode has changed from paper to computer.

The present paper presents the main aims of a large-scale diagnostic assessment project launched at the University of Szeged in Hungary in 2009 and shows the result of the first media effect study carried out in this project.

Objectives

The long-term project aims to design an online formative assessment system for the first six grades of primary school. The objective of the first phase is to pilot the system in 150 schools and study the related technological and methodological issues in detail.

In this paper we (1) outline the formative (diagnostic) assessment system; (2) present the results of the first two years of the project, in which online testing was introduced and piloted in various age groups in different school subjects; and (3) compare results of paper-and-pencil (PPT) and online testing in order to identify domains and item formats where the two media may influence achievements.

While high-stakes Computer-Based Assessment (CBA) requires stricter technological conditions (e.g., standard equipment, security, and confidentiality),

formative testing is easier to introduce by using schools' existing equipment and can be adapted to local circumstances.

A step-by step introduction of the new features offered by CBA, the approach the project applies, allows us to control media effect. A detailed comparison and analysis of PPT and CBA establishes scientific bases to improve validity of computerized tests.

The Diagnostic Assessment Project

In most of the developed countries international and national assessment programs provide comprehensive feedback on trends in students' achievements regularly. However, they are not suitable for tracking students' individual development, diagnosing learning difficulties or identifying causes of failure, or supporting different solutions. Fostering students' learning processes and facilitating their development require other types of information as well as frequent more accurate and detailed personal feedback.

These requirements have resulted in the launch of significant R&D programs worldwide. Projects aiming at addressing individual student needs by directly supporting learning and instruction and developing diagnostic measurement systems are being conducted in several countries. These measurement systems must rely on scientifically elaborated theoretical frameworks and empirically tested standards. Diagnostic assessments can only be adequately conducted with the help of modern information technology devices, keeping the required frequency and accuracy. Such devices are also needed for fast automatized processing and analysis of data as well as for the provision of accurate feedback.

The „Diagnostic assessments" research and development program of the Center for Research on Learning and Instruction, University of Szeged, will lay the foundations for such a nationwide system in Hungary. The activities of the project launched in 2009 are organized into seven smaller projects:

(1) Devising assessment frameworks for reading, mathematics and science for the first six grades of primary school;
(2) Exploring diagnostic assessments in further cognitive and affective domains (Social skills, English as a foreign language, Visual skills, Civic education, Motivation, Health literacy and health behavior and Learning to learn);
(3) Developing item banks in reading, mathematics and science (ca. 3 x 600 items); the full-scale administration of the items in PP mode is administered in this module to determine PP item parameters;
(4) Creating a platform for online testing by adapting TAO and migrating a selection of items into TAO in 2010; furthermore, a pilot TBA will be implemented in the schools equipped well enough in spring and autumn 2010 and spring 2011;
(5) In-service training of teachers to prepare them to use the system;
(6) Devising diagnostic assessment instruments for students with special education needs (SEN) students by making „unimpeded" and easier tasks with the assistance of teachers for handicapped children and developing special computer interfaces for SEN students;
(7) Meta-analysis of the data of national and international assessments (TIMSS, PIRLS, PISA).

An open source assessment platform, TAO[1], is used as a delivery platform, and in the first phase of the project the online assessment system will be piloted in ca. 200 partner schools.

The research design of the project allows us to devise a large number of items both on paper and on computer to build PP and CB item banks and to compare the achievements and item parameters on the tests using different media.

Methods

Participants and assessment instruments

The present analyses synthesize the results of six PP and six CB data collections (Table 1) in the first two years of piloting and implementation of CBA in Hungary. In the first year, a nationally representative sample of 843 5th grade students from 34 schools was tested on an inductive reasoning test. Identical versions of the test were administered in paper and online formats. Because of the abstract content of the reasoning test, no learning took place from testing; therefore, the same items could be delivered twice in two media.

The test comprises three subtests: number analogies (14 open ended items), number series (16 open ended items) and verbal analogies (28 multiple choice items).

Table 1. Sample, field, and research design of the studies

Year	Grade	N (CB)	Field	Mode of administration	Design
2008	5	843	Inductive reasoning	PP -> CB	same sample
2009	2	510 (96)	Inductive reasoning	PP -> CB	different samples (anchor items)
2009	2	285	Reading	PP -> CB	same sample
2009	6	449	Reading	PP -> CB CB -> PP	same sample
2009	6	58	Mathematics	PP -> CB CB -> PP	same sample
2009	6	598 (189)	Problem solving	PP -> CB	different sample

In the second year, in 2009, online testing took place in the same schools; this time the samples were drawn from 2nd and 6th grade students (see Table 1). Four different tests were administered: inductive reasoning, reading, mathematics, and problem solving. Reading comprehension was delivered to both cohorts. Electronic reading and problem solving skills are obviously related to the modern digital world, and technology offers excellent means to assess them; however, in the first phase of the project we did not deal with these specific issues. The same content was covered; however, different items were used for the two media.

[1] https://www.tao.lu/

The inductive reasoning test was developed directly for young learners (Csapo, 1997). Due to the young age of the target population, special attention was paid to both the verbal for the non-verbal character of the test; one of the subtests contained many pictures, figures and images and as little reading text as possible in order to avoid measuring students' reading skills instead of their inductive reasoning skills. The two other subtest contained anchor items to the inductive reasoning test used by the senior cohort. Regarding the item types, the test comprised both open-ended and multiple-choice items.

The reading comprehension tests consisted of two subtests: a continuous (a narrative) and a non-continuous text-type (maps, diagrams, tables). The continuous text subtest in both formats contained a description, and the context of the PP and CB texts was personal in line with the students' age. The non-continuous texts' content situation was educational and the type of these texts was document. The 36 item tests constructed for grade 6 comprised dichotomous and multiple choice questions, while the 16 item test constructed for grade 2 included multiple choice and true or false items. The PP and CB tests in both cohorts were parallel test versions; they included the same instructions, response types and number of items.

The mathematics test was a criterion-referenced test and measured the knowledge of 6 grade learners. The test comprised multiple-choice items and short-answer items in the two media. Equivalent test versions were created for measuring the media effect with some identical items across testing media.

Finally, the problem solving test comprised short and longer texts, tables, and pictures. All tasks were embedded in a single authentic situation of a family trip. The pages of the text booklet were divided into two columns. The left column presented information in realistic formats (e.g., map, picture, drawing, and table), whereas the right column described the story of a family and prompted students to solve problems as they appear during their travel.

Multiple-choice and open-ended items required 1-2 character-long answers, layout on paper and screen was kept as similar as possible; the only difference concerned the delivery media. Test takers are expected to answer open-ended items by writing down in PP mode and typing via keyboard in CB mode, whereas they are to circle the appropriate letter in multiple choice items in PP format and to use a radio button in CB format. PP items requiring long answers especially in case of problem solving were converted into multiple choice items by CB testing (see Table 2). The tests contained anchor items allowing Rasch analysis of the achievements on the same scale.

Table 2. Different item types of the tests

Study	Item types (PP)	Item types (CB)
2_ind	McQ, SA	McQ, SA
2_read	McQ, SA, T/F	McQ, TF
5_ind	McQ, SA	McQ, SA
6_read	T/F, McQ	T/F, McQ
6_math	McQ, SA	McQ, SA
6_ps	McQ, SA, LA	McQ

T/F: true or false; McQ: multiple choice; SA: open ended and requires short answer (1-2 character); LA: open ended and requires long answer

Procedure

Computers available at the participating schools were used for the online assessment. Students used the operation systems and browsers installed on the computer. No special effort was devoted to standardizing the equipment.

Each participant took the cognitive test in 5th grade, the reading comprehension test in 2nd and 6th grade and the mathematics test in 6th grade in both PP and CB mode. In the case of inductive reasoning test in 2nd grade and problem solving test in 6th grade different sample were used in the PP and CB data collection.

Besides of reading comprehension and mathematics test in grade 6, in all cases the PP tests was taken first in regular classrooms, then, a few weeks later the online version of the tests was taken in specially equipped computer rooms. The tests were time-limited; participants had 35 minutes to complete each test in both formats, separately. The computer-based version of the test was delivered via the Internet (see Csapó, Molnár, & R. Tóth, 2009) with the TAO (Testing Assisté par Ordinatur – Computer-Based Testing) platform. 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 (Plichart, Jadoul, Vandenabeele & Latour, 2004).

Dichotomous data were used in the analyses. Detailed item analyses were performed by using several means of classical test theory and IRT. The first test-level characteristic that is of importance is test reliability. Cronbach's alpha was used for analyzing the reliability of the tests and subtests in both formats. To describe the achievement differences in PP and CB format, test scores t-test was computed. By analysing item-level differences according to the media item difficulty parameters were computed and compared.

Results and discussion

Reliability

The reliability indexes of the PP and CB assessments in all fields are compared in Table 3. Data show that the reliability indexes of the PP inductive reasoning test in Grade 2 differ (α=.87 and .80, respectively) but in Grade 6 did not differ significantly (α=.91) from the reliability of the CB inductive reasoning test (α=.90). The differences can be caused by the change of some items, where the same items were used in CB and PP format, the reliability indexes did not change via the media.

In the field of reading comprehension and problem solving, where the items contain texts and students need to read more to answer the questions and solve the problems, the reliability indexes change according to the media. In CB environment the reliability indexes are lower than in PP testing, whereas in the case of inductive reasoning and mathematic literacy the test mode-mode effect was not noticeable regarding measured reliability indexes.

Table 3. The reliability indexes in PP and CB mode

Grade	Test	Reliability (PP)	Reliability (CB)
2	Inductive reasoning	.87	.80
	Reading comprehen.	.91	.74
5	Inductive reasoning	.91	.90
6	Reading comprehen.	.76	.70
	Mathematic literacy	.95	.95
	Problem solving	.78	.58

Test and subtest-level mean achievement differences

The test and subtest level analyses indicated different results regarding context, item type and age (Table 4). Comparing students' mean achievements on PP and CB-based testing, there was no media effect on tests tapping into reasoning through an abstract content. The finding was independent of the targeted cohort.

A different picture can be seen in the field of reading comprehension and problem solving depending on the targeted cohort. In Grade 2 there was no significant achievement difference; students' achievement was free of test mode, while in grade 5 and 6 the average scores on the PP test and the online test-version differed significantly (see Table 4 and 5). On the reading test students' achievement was higher in PP format than in CB format, while students solved the problems more effectively in the online environment than in the PP format.

Table 4. Test-level achievement differences

Grade	Test	PP (%)	CB (%)	Sig.
2	Inductive reasoning	36.7	40.6	n.s.
	Reading comprehension	78.2	76.5	n.s.
5	Inductive reasoning	27.2	26.0	n.s.
6	Reading comprehension	74.9	69.2	PP>CB
	Mathematic literacy	53.3	41.5	PP>CB
	Problem solving	23.9	34.7	PP<CB

A similar picture was found in the case of mathematic literacy, where students outperformed their results in PP over CB mode.

Table 5. Subtest-level achievement differences

	Grade	Subtest	PP (%)	CB (%)	Sign.
Inductive reasoning	2	Non-verbal	36.7	40.6	n.s.
	2	Verbal analogy.	-	24.8	-
	5	Verbal analogy	40.3	42.8	PP<CB
	2	Number analogy	-	15.6	-
	5	Number analogy	27.0	25.4	PP>CB
	5	Number series	14.3	9.9	PP>CB
Reading comprehension	2	Continuous text	85.4	69.8	PP>CB
	6	Continuous text	68.4	60.5	PP>CB
	2	Non-continuos text	66.6	81.5	PP<CB
	6	Non-continuos text	77.8	72.9	PP>CB
Problem solving	6	Table, picture, short text	31.8	40.2	PP<CB
	6	Table, picture, longer text	22.4	24.3	n.s.
	6	Picture	29.2	36.6	PP<CB
Mathematics	6	-	53.3	41.5	PP>CB

All in all, at the beginning of the elementary school independently of the measured area and context, there were no differences in achievements between students' PP and CB test results. However, at the age of 12 irrespective of the measured area, achievements were different in the PP and CB test results. Students' achievements were higher in PP mode on tests more closely connected to school subjects, whereas their performances were higher in CB format on tests connected to real life. This result suggests TAP (transfer appropriate processing; Clariana, & Wallace, 2002) as a possible explanation of the test mode effect. According to TAP, lesson and assessment processes should correspond. If the learning process and activities are PP oriented, student achieve better in PP environment than on the computer.

The subtest level achievement differences give a more detailed picture about the media effect. Subtest-level analyses indicated significant differences in students' performance on most of the test-types.

In principle students performed better on the subtests containing tasks requiring more calculating (see Table 5, inductive reasoning: number series, and analogies and mathematics test) or reading long passages on printed medium, while they had fewer difficulties on non-continuous texts (comparing maps, charts, diagrams) or information retrieving from tables on screen presentation tests. This finding corresponds to the results in connection with the test mean achievements and the TAP explanation.

Item-level analyses

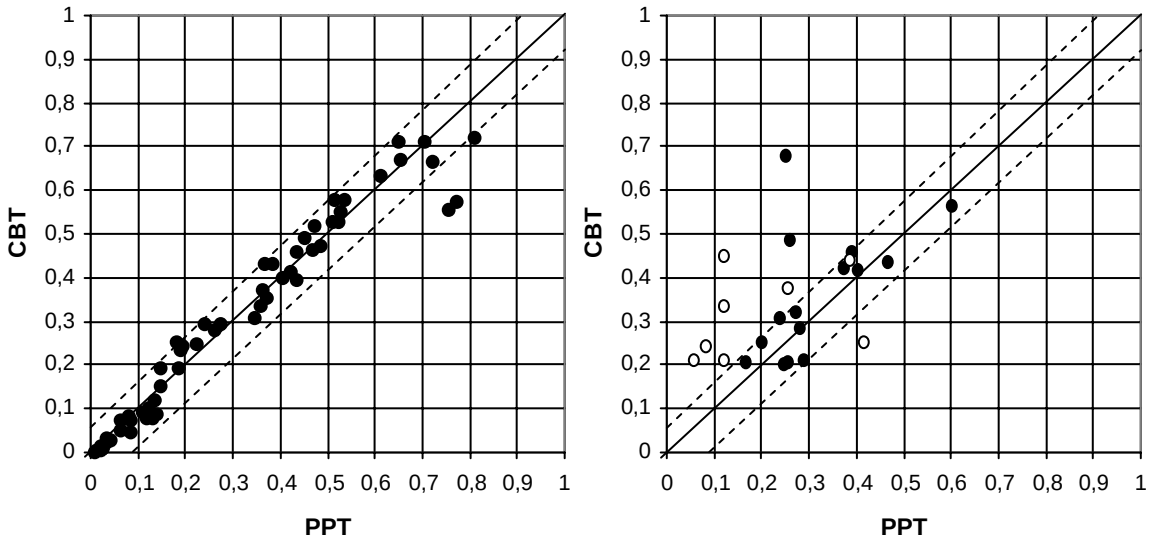
The item-level analyses indicate that

(1) multiple choice items involving a small amount of reading do not change the difficulty level; thus, there is a minimal medium effect;
(2) multiple choice items involving a large amount of reading increase the difficulty level; a higher medium effect is to observed;
(3) the highest medium effect was noticeable in the case of items having different item types (open ended to McQ), most probably the difference is in these cases the role of item type and not the impact of medium.

Figure 1 shows the item level comparison of the inductive reasoning test (grade 5) and the problem solving test (grade 6). The black signs represent items with the same item type; whereas white signs represent items whose item type was changed by migrating the items from PP to CB format (open ended items were converted to McQ items).

According to the above tendencies, the difficulty level of the multiple choice items requiring a small amount of reading did not differ significantly in PP and CB format (black signs). The highest media-effect was noticeable on items whose types were different in CB and PP mode (white signs).

Figure 1. Item difficulty level in PP and CB mode (black: the same item type, white: open ended item to McQ item)



Inductive reasoning (Grade 5) Problem solving (Grade 6)

Influencing factors

Previous research has suggested that computer familiarity, gender, and competitiveness are learner characteristics that relate to test mode effect (e.g., Clariana, & Wallace, 2002). In this present investigation only the gender variable was available in

all of the studies and did not prove to be a impact factor in the mode-effect analyses. Boys and girls achieved the same way independently of the applied medium.

Conclusion

As the data collections in this project took place in average schools by using their actual infrastructure, the pilot studies have indicated that even these equipment not standardized for assessments generate reliable results. In the long run, the system developed in the framework of the project will be used for low stakes assessments and detailed frequent student-level feedback, and in this context, the studies suggest that technology can be used for this purpose without major difficulties.

Although the computerized assessments took place only in a limited number of domains, the assessment instruments represented a variety of content. The overall results indicate that the media significantly affect the performances. On the other hand, the differences were not large at the test level, and in most cases the causes of the differences could be identified. The results suggest that if the goal was to develop equivalent summative assessment for the two media, it could be achieved by careful analyses and modification of the items. However, if the goal is to develop CBA instruments for individual frameworks, a more important issue is the correspondence between the postulated constructs and the items used for their assessment. In this context the equivalence of the two media is not a concern but studying the particular differences between the two media may support a developmental process towards the improvement of the validity of online assessment.

The methods developed in the pilot phase can be applied both for analyzing the equivalence of paper-and-pencil and online tests and for developing scales on both media. Further research is needed to study the effect of media at other domains, to identify the differences of the cognitive processes relevant in the two media and for controlling the effects of other variables that were not the goal of these studies (e.g., different interest, motivation and attitudes towards the different media, the role of the actual hardware and software).

References

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B, Kaplan, B., & Yan, F. (2008): Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment*, 6. No. 9. http://escholarship.bc.edu/jtla/vol6/9/

Clariana, R. & Wallance, P. (2002): Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, 33. No. 5. 593-602.

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., Molnár Gy., & R. Tóth, K. (2009). Comparing paper-and-pencil and online assessment of reasoning skills. A pilot study for introducing electronic testing in large-scale assessment in Hungary. In F. Scheuermann, & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 113-118). Luxemburg: Office for Official Publications of the European Communities.

Csapó, B., Latour, T., Bennett, R., Ainley, J., & Law, N. (2009): Technological Issues of Computer-Based Assessment of 21st Century Skills. Draft white paper. http://www.atc21s.org/GetAssets.axd?FilePath=/Assets/Files/dc7c5be7-0b3a-4b7d-8408-cc610800cc76.pdf

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. No. 2. http://escholarship.bc.edu/jtla/vol5/2/

Plichart, P., Jadoul, R., Vandenabeele, L. és Latour, T. (2004): TAO, a Collective distributed computer-based assessment framework built on semantic web standards. In Proceedings of the International Conference on Advances in Intelligent Systems – Theory and Application AISTA2004, In cooperation with IEEE Computer Society, November 15-18, 2004. Luxembourg, Luxembourg.

Wang, S., Jiao, H., Young, M., Brooks, T., & Olson, J. (2007): A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, 67. No. 2. 219-238.

Wang, S., Jiao, H., Young, M., Brooks, T., & Olson, J. (2008): Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, 68. No. 1. 5-24.

Acknowledgements

The first phase of data collection took place in the framework of the *Hungarian Educational Longitudinal Program* and carried out by the *Research Group on the Development of Competencies, Hungarian Academy of Sciences* (MTA-SZTE Képességkutató Csoport).

The *Diagnostic Assessments Project* is supported by *Hungarian Development Agency* (TÁMOP 3.1.9).