

Comparing Paper-Based and Computer-Based Testing in the First Grade

Paper presented at the Round Table on
Uncovering the Promise and Pitfalls of Computerized and Adaptive Testing in Action,
AERA, 2011, New Orleans, Louisiana, USA

Gyongyver Molnar, Krisztina R. Toth and Beno Csapo
University of Szeged

Abstract

Using computers for administering tests opens new possibilities, but poses new questions. Using CB testing by very young students poses several challenges; therefore special attention must be paid for controlling validity and reliability of the instruments. The purpose of this paper is to study the media effects in a curriculum-independent competency field by first grade (age 6) students. Due to the young age of the target population, the instrument of the study consisted figural, nonverbal items. The same inductive reasoning test was administered in PP and in CB mode. Student-level differences indicated media effect by first grade student that is independent of construct measured and item type applied. First graders achievement was higher in PP testing than in CB format.

In the world, almost everywhere one looks technology abounds. Within the last few years, these tools and applications have fundamentally changed the way people live, learn and communicate. There is no longer doubt that multimedia application design offers new insights into the learning process, gives possibilities to represent information and knowledge in a new and innovative way and have the potential to transform education. However, technology is not the issue; it is a catalyst and provides new opportunities to improve the quality of education including educational assessment.

Technology-based assessment (TBA) has many advantages (Bridgeman, 2009) and possibilities. According to the international tendencies (e.g. OECD, ETS, NCES) in educational assessment the use of TBA is increasing. Major international projects focus on implementing TBA (see e.g. Assessment and Teaching of 21st Century Skills by Intel, Microsoft, and Cisco Education Taskforce, 2008; PISA 2012 Complex Problem Solving by OECD). It is without doubt that TBA will replace paper-based testing, and extend business and substance (Bennett, 2002) of assessment in education. Parallel to these tendencies paper-based assessment reached its limits (Scheuermann, & Björnsson, 2009). Further development (e.g. Reduction of costs, logistic and feedback time) is unexecutable with traditional – paper-based – assessment tools (Molnar, 2010).

TBA (1) opens new areas (e.g. dynamic assessment), enables measuring new constructs (e.g. problem solving in rich technology environments, Bennett, Persky, Weiss & Jenkins, 2007) and assessing dynamic, where testee faces a dynamically changing

environment (Greiff and Funke, 2009). TBA (2) rises new issues in assessment (e.g. educational data mining – log file analyses; eye and face tracking) and (3) offers new assessment methods (e.g. adaptive testing – see Frey, 2007), that cannot be realized otherwise. In case of an adaptive test the difficulty of the test tailors dynamically to the student's ability level so that subsequent items are selected from an item bank dependent at a difficulty appropriate for the student. This provides more time-efficient and accurate assessments. TBA (4) increases motivation (e.g. task adapts to the examinee's ability level) and (5) changes the whole assessment process including item generation, scoring, data-processing, information flow, feedback and the speed of assessment. It provides rapid and precise feedback for the participants and stakeholders that cannot be achieved by paper-based testing. Finally TBA (6) poses new questions and problems, for example validity issues regarding media effect studies when TBA is applied to replace traditional paper-based assessment and when skills related to the digital world are assessed (Csapó, Latour, Bennett, Ainley, & Law, 2009).

Validity issues regarding media-effect studies belong to one of the key research areas. The transition from PP to TBA in educational context requires a step-by step procedure (Csapo, Molnar, & R. Toth, 2009); the first is the adequate control of media effect during testing and make detailed comparisons of PP and CB test results. 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. Several studies have been conducted with older students and adults to evaluate the comparability of CB and PP scores or to measure the effect of administration mode (e.g. see Clariana and Wallance, 2002; 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; Csapo, Molnar, & R. Toth, 2009). However, only a few ones focused on testing in early childhood education in CB environment, and most of them used mathematical context (Choi and Tinkler, 2002).

For our study of media effect a general cognitive ability field was chosen. Klauer's theory of inductive reasoning formed the ground of devising our test, as he constructed an elaborated system of inductive reasoning by defining its elements and their relationships (Klauer, 1989; Klauer and Phye, 2008).

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 early childhood.

Educational and scientific importance

Recent tendencies in large-scale international educational assessment programs (e.g. OECD, ETS, NCES) indicate that CB testing plays an increasing role. Future European surveys plan to introduce computer-based assessments of student achievements, therefore, students are expected to be familiar with CB testing. This fact may facilitate the early use of CB testing at school. However, most cases studies do not focus on pupils' CB testing in an educational context.

In connection with media effect control, the detailed analyses of a comparison of PP and CB testing establish scientific bases to improve the efficiency, effectiveness and validity of computerized tests.

Objectives

A long-term project in Hungary aims at implementing an online diagnostic assessment system for the first six grades of primary school. Testing very young students (especially in grade 1 and 2, at the age of 6-8) poses several challenges; therefore, special attention must be paid to the validity and reliability of the instruments. The purpose of this paper is to study the media effects in a cross curricular competency field by first grade (age 6-7) students, to control media effect and make detailed comparisons of test results delivered by different media.

In this paper we

- (1) outline the formative (diagnostic) assessment system;
- (2) present how first grade students solve computer-based tests;
- (3) compare their achievement in PP and CB mode;
- (4) identify the item formats where the two media may affect the achievements; and
- (5) compare sub-groups of the samples to characterise students who achieve better or worse in CB tests.

The Diagnostic Assessment Project

At present, national and international assessment programs are utilized on a regular basis to provide comprehensive feedback on the achievement trends of Hungarian students. They are not, however, suitable for tracking the individual development of students, diagnosing learning difficulties and/or identifying causes of failure. Fostering students' learning processes and facilitating their development require other types of information as well as frequent more accurate and detailed personal feedback. Diagnostic assessments must exploit the full range of possibilities that modern information technology offers in order to be ultimately effective. Exploitation of advanced technology ensures that frequency and accuracy concerns are properly addressed. In addition, the efficiency with which modern technology can process and analyze data adds the benefit of access to immediate feedback.

Advancement in the fields of cognitive research and educational evaluation has provided the essential theoretical knowledge necessary for development of new and more sophisticated diagnostic assessment frameworks. Large-scale national and international programs have renewed interest in the psychometric basis of assessment. In addition, the rapid development of information-communication technology has contributed essential innovations needed for computer-based assessment. Despite these huge technological advances, a copious amount of work is still required to utilize and organize these innovations into a workable online diagnostic assessment system.

Supported by The Social Renewal Operational Program, the Developing Diagnostic Assessments research and development project of the Center for Research on Learning and Instruction, University of Szeged, is laying the foundation for a nationwide diagnostic assessment system for the first six grades of primary school.

The project involves

- 1) the development of assessment frameworks;
- 2) exploration of the possibility of using diagnostic assessments in various fields;
- 3) item construction and the development of item banks;
- 4) the setting up and piloting of an online assessment system;
- 5) extension of assessment to students with special educational needs;
- 6) preparation of teachers and educational experts for participation in a variety of assessment processes; and

7) secondary analysis of data collected in national and international assessment programs.

The objectives of the project are realized through the efforts of a team of interrelated and collaborative working groups formed expressly for the development and execution of the project. The developmental work is achieved through national and international cooperation with the goal of attaining the best expertise and knowledge available to inform the project.

Methods




Participants and assessment instruments

A nationally representative sample of 1st grade students were tested with a paper-based inductive reasoning test in the spring of 2008, around the end of the school year (age 6-7; N=5156). In 2010, an online testing took place with the computerized version of the same inductive reasoning test. The sample for the study were drawn from 1st and 2nd grade students (N=313).

Because of the abstract content, an inductive reasoning test was developed directly for young learners. Due to the young age of the target population, special attention was paid to the non-verbal character of the test, i.e. it had to contain 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 (see Figure 1). It consisted of 37 figural, non-verbal items. Regarding the item types, the test comprised both open-ended and multiple-choice items.

The structure of the test is based on Klauer's definition (1989, 2008) of inductive reasoning, i.e. the items belonged to the six sub-classes of inductive reasoning (generalization, discrimination, cross-classification, recognizing relations, discriminating relations, system formation).

Figure 1. Examples of tasks in the inductive reasoning test with the measured classes of inductive reasoning

Generalization	Underline those 3 items which have one feature in common that the other two do not have.	
Discrimination	Underline that one which does not fit with the others.	
Discrimination relations	Underline the item which disturbs the given order.	

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.

The online data collection was carried out with the TAO (Testing Assisté par Ordinateur – Computer-Based Testing) platform via Internet (see Csapo, Molnar, & R. Toth, 2009). 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).

The same test was used in paper-pencil and computer-based format. The paper and screen layout were kept as similar as possible, only the delivery media was different.

Detailed test, subtest and item-level 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.

Analysis and results

Test and subtest-level mean achievement differences

The reliability index of the PP inductive reasoning test (Cronbach- α =.88) did not differ significantly from the reliability of the CB inductive reasoning test (Cronbach- α =.85).

Student-level differences indicated that in contrast with previous media effect studies in the same field focusing on higher grade students (see Csapo, Molnar, R. Toth, 2009); there is a media effect noticeable at first grade students. First graders' achievement was higher in PP (M=45.33%, sd=20.07%) than in CB format (32.66%, SD=18.17%; t=6.11, p<.01). The media effect resulted in an amount of one year of difference, students in the first grade achieved on PP test like second graders (x=47.71%, sd=18.04%) in an online environment (see table 1).

Table 1. Means and standard deviations of the test for inductive reasoning in CB and PP mode (%)

Grade	PP test		CB test		t-test	
	M	SD	M	SD	t	p
1	45.3	20.1	32.7	18.2	6.11	p<.01
2	-	-	47.7	18.0	-	-

The subtest level analysis gives a more detailed picture on the media effect. Regarding the basic structures of inductive reasoning, first graders achieved at the same

level both in PP and CB subtest than on the whole one. It indicates that regardless of the construct measured and the item type applied, in case of first graders the delivery media caused significant achievement differences. However, depending on the used item format the amount of differences are changing. The highest differences were found in the field of system formation, while the lowest in cross-classification (see table 2).

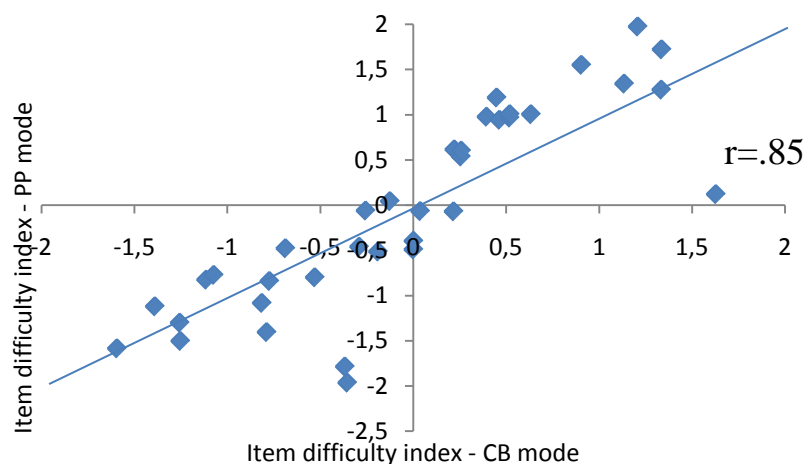
Table 2. Means and standard deviations of the six subtests for inductive reasoning in CB and PP mode (%)

Subtest	PP test		CB test		Diff.
	M	SD	M	SD	
Generalization	46.1	27.8	34.3	24.0	11.8
Discrimination	52.7	26.0	42.1	25.9	10.6
Cross-classification	26.2	27.4	20.9	20.3	5.3
Recognizing relations	48.4	25.9	33.9	25.0	14.5
Discrimination relations	36.8	25.9	30.0	29.5	6.8
System formation	49.7	31.0	30.1	26.8	19.6

Item-level analyses

Figure 2 shows the item level comparison of the inductive reasoning test in computer-based and paper-based mode. According to the above tendencies, the difficulty level of the items proved to be easier in PP mode than in CB mode. The highest media effect was noticeable at items, where the answer and distractors contained more figures. The influential factor of the delivery media proved to be the strongest ($t=8.24$, $p<.01$) at these items. In this case, pupils actually have more information available and need to take into consideration more data on the screen at the same time.

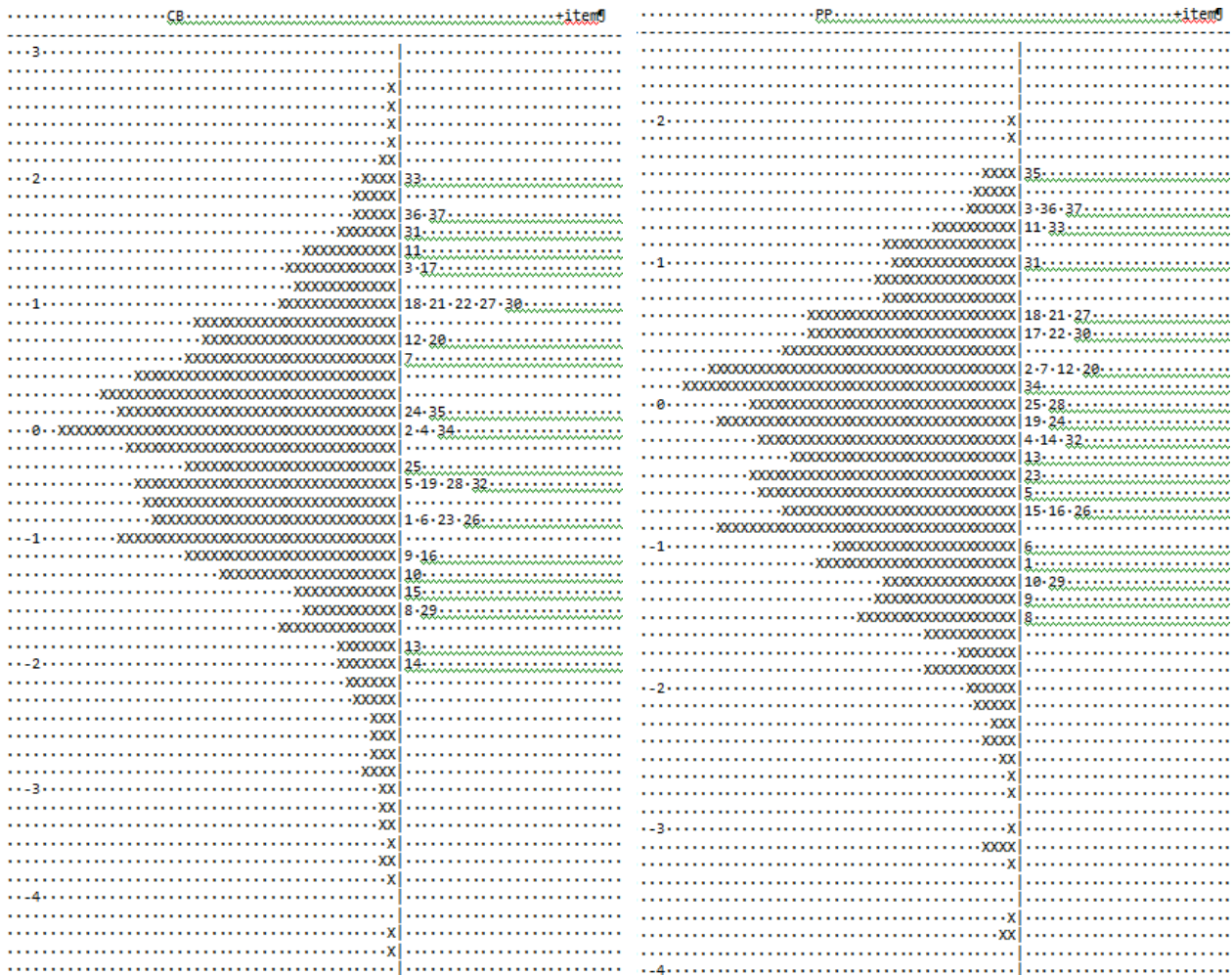
Figure 2. Item difficulty level in PP and CB mode



The correlation between the item difficulty levels in PP and CB mode is significant ($r=.85, p<.01$).

The similarity of the item/person maps indicates the analogous behavior of the test in CB and PP mode (Figure 3). The logit scales of the item/person maps are directly not to compare, but the relation of the items and the shape of the distribution curves indicates similar behavior of the test in the two media.

Figure 3. Item/person map of the test in CB and in PP mode



Influencing factors

Regarding gender analyses, there were no achievement differences between the achievement of boys and girls in CB test results. Similar result was found at subtest level as well. Students' socio-economic factors (e.g. number of books, number of PCs, number of mobile phones) did not influence their CB test scores; similarly the common usage of computer and/or internet did not result in higher CB test results (see table 3).

The delivery media had a significant impact on first graders' achievement regardless of pupils' different background variable the measured thinking structure, context and item format.

Table 3. Means and standard deviations of the CB test in gender division (%)

Grade	Gender	CB test		t-test	
		M	SD	t	p
1	male	31.2	17.9	-.84	n.s.
	female	34.2	18.3		
2	male	47.1	18.0	-.67	n.s.
	female	48.5	15.8		

Conclusion

The data collection in this project took place in average schools by using their actual infrastructure. The pilot study has indicated that even these equipment not standardized for assessments generate reliable results and fits for the introduction of computer-based assessment even in early childhood education. 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.

There exist different methods to study the effect of the administration mode in educational testing: to compare results on test, subtest and item level; to compare achievements according to the context, type and complexity of the items and to identify factors that influence student behaviour dependent on the delivery media. To make the above mentioned comparisons more accurate, and lend more support to the media studies, in this study – against the international tendencies– we focused on testing in early childhood education.

The overall results indicate that the media significantly affect the performances even at the age of 6 to 8. First grade students did not have any trouble to solve computer-based tests. The differences were larger at the test level, than measured in studies conducted with older students in the same field (Csapo, Molnar & R. Toth, 2009, 2010). The level of difference corresponds to one year of development; pupils in the first grade achieved in PP environment like second graders in CB test. This finding is in contrast with previous media effect studies in the same field focusing on higher grade students (see Csapo, Molnar, R. Toth, 2009); where no achievement differences were detected.

Regardless of the item type applied the delivery media caused significant achievement differences. However, depending on the used item format the amount of differences are changing. The highest media effect was noticeable at items where the answer and distractors contained more figures. In this case, pupils actually have more information available and need to take into consideration more data on the screen at the same time.

According to the validity issues, the comparison of PP and CB test results indicated no differences in the reliability indexes of the same test in PP and CB mode. The data are on the same level generalisable, independent of the delivery medium.

Regarding gender analyses, there were no achievement differences between the achievement of boys and girls in CB test results. Similar result was found at subtest level as well. Students' socio-economic factors did not influence their CB test scores; similarly the common usage of computer and/or internet did not result in higher CB test results.

The delivery media had a significant impact on first graders' achievement regardless of pupils' different background variable the measured cognitive structure, context and item format. The results suggest that if the goal was to develop equivalent summative assessment

for the two media, studying the particular differences between the two media using different research condition and research design may support a developmental process towards the improvement of the validity of online assessment. Further research is needed to study the effect of media in 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.

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 Képeségkutató Csoport).

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

Gyongyver Molnár was having Bolyai Janos Research Scholarship at the time of writing the present paper.

References

- Bennett, R. E. (2002). Inexorable and Inevitable: The Continuing Story of Technology and Assessment. *Journal of Technology, Learning and Assessment*, 1. 1. sz. <http://escholarship.bc.edu/jtla/vol1/1/>
- 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/>
- 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.
- Bridgeman, B. (2009). Experiences from Large-Scale Computer-Based Testing in the USA. In F. Scheuermann, & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 39-44). Luxemburg: Office for Official Publications of the European Communities.
- Choi, S. W., & Tinkler, T. (2002). Evaluating comparability of paper-and-pencil and computerbased assessment in a K-12 setting. Oregon Department of Education. <http://www.ncme.org/repository/incoming/100.pdf>
- 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.
- 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>

- Csapo, B., Molnar, Gy., & R. Toth, 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 Scheuermann, F. & Björnsson, J. (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.
- Csapo, B., Molnar, Gy., & R. Toth, K. (2010). *Implementing an Online Formative Assessment System: From Paper- Based to Computer-Based Testing*. AREA, Denver, USA, April 29- May 4. 2010. p. 213.
- Frey, A. (2007). *Adaptives Testen*. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 261-278). Berlin, Heidelberg: Springer.
- Greiff, S. & Funke, J. (2009). On the way to competence levels in dynamic microsystems: The MicroDYN Approach. In J. Funke, J. Wirth & S. Greiff (Eds.), *Symposium on Problem Solving. Assessment of Problem Solving Competencies*. Paper presented at the EARLI in Amsterdam, The Netherlands, 25.08.-29.08.2009.
- 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/>
- Klauer, K. J. & Phye, G. D. (2008). Inductive reasoning. A training approach. *Review of Educational Research*, 78, 85-123.
- Klauer, K. J. (1989). *Denktraining für Kinder I*. Göttingen: Hogrefe.
- Molnar, G. (2010). *Technology-based assessment: Challenges and Promises*. Keynote presentation, EDEN, Budapest, Hungary. 24-27, October, 2010.
- 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.
- Scheuermann, F. & Björnsson, J. (2009, Eds.). *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*. Luxemburg: Office for Official Publications of the European Communities.
- 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.