Summary of PhD dissertation

KRISZTINA TÓTH

THE COMPARATIVE ANALYSIS OF PAPER-AND-PENCIL AND COMPUTER-BASED INDUCTIVE REASONING, PROBLEM SOLVING AND READING COMPREHENSION TEST RESULTS OF UPPER ELEMENTARY SCHOOL STUDENTS

Supervisor
Dr. Gyöngyvér Molnár
habilitated associate professor

University of Szeged
Faculty of Arts
Doctoral School of Education
Doctoral Program of Information and Communication Technologies in Education

Szeged
2015

INTRODUCTION

In the past four decades several institutes dealing with educational measurement and assessment have started to adapt their paper-and-pencil tests into computerized environment (*Csapó, Ainley, Bennett, Latour & Law*, 2012). Several factors motivate the transfer of paper-and-pencil tests to an online environment (*Bennett*, 2003; *Bodmann & Robinson*, 2004; *Choi & Tinkler*, 2002; *Csapó*, *Molnár & Tóth*, 2008; *Horne*, 2007), for example the simplification of administration, the long-term cost efficiency, the possibilities provided by the computer-based measurement and assessment such as new item types (*Tóth*, *Molnár*, *Wüstenberg*, *Greiff & Csapó*, 2011), the use of test algorithms (*Scalise & Gifford*, 2006; *Schröders & Wilhelm*, 2011), and log-file analysis (*Goldhammer*, *Naumann*, *Stelter*, *Tóth*, *Rölke & Klieme*, 2014; *Tóth*, *Rölke & Goldhammer*, 2012a, 2012b; *Tóth*, *Rölke*, *Greiff & Wüstenberg*, 2014; *Tóth*, *Rölke*, *Naumann & Goldhammer*, 2012; *Tóth*, *Wüstenberg*, *Rölke & Greiff*, 2012).

Even though computer-based testing has been around in several forms for many decades (*Asuni*, 2009) and it is applied in more and more areas, it makes educational researchers face several challenges (*Paek*, 2005) even at the level of presenting their paper-and-pencil tests in an online environment. With the introduction of a new testing environment some research questions arise. The first question is whether the change of test environment exerts an influence on students' achievement and the second question is whether it is assured that the learner is not advantaged or disadvantaged by using a computer-based instrument instead of a paper-and pencil one (*Paek*, 2005). No definite answer can be given to these research questions (1) because technology is constantly developing, thus technological and test interface parameters are changing and (2) because of individual differences of learners participating in the testing and (3) because the assessed domain and the assessment instrument might influence students' achievement. Therefore, when data are derived from two different testing environments, they must be interchangeable in order to draw scientific conclusions.

THEORETICAL BACKGROUND

Several international guidelines deal with the equivalence of results achieved at the paper-and-pencil and electronically administered tests (*Bartram & Coyne*, 2005). The equivalence of test scores achieved in paper-and-pencil and computerized environments needs to be justified in an empirical way (*Paek*, 2005). If the same test taken in paper-and-pencil or computerized environment does not result in equivalent student achievements (*Leeson*, 2006), the mediating tool affects test results. This phenomenon is further referred to as the mode effect (*Clariana & Wallace*, 2002). In case the equivalence of the test and the test results are assured, students' achievements are test mode independent.

Research on the influence exerted by the testing environment on achievement has been at issue since the 1970s (*Gaskill & Marshall*, 2007). Thus, the examination of the achievement differences caused by the testing environment and the identification of the underlying causes have been in the scope of more than 300 international studies carried out and published in the past decade (*Wang*, *Jiao*, *Young*, *Brooks & Olson*, 2008). Some studies focusing on the examination of mode effect do not find significant differences between the results achieved in paper or online (*Bodmann & Robinson*, 2004; *Horkay*, *Bennett*, *Allen*,

Kaplan & Yan, 2006; *Johnson & Green*, 2006; *Poggio*, *Glasnapp*, *Yang & Poggio*, 2005; *Puhan*, *Boughton & Kim*, 2007; *Wang* et al., 2007, 2008). On the other hand, other mode effect studies justify significant differences between the achievements in the two testing environments (*Bennett*, *Braswell*, *Oranje*, *Sandene*, *Kaplan & Yan*, 2008; *Csapó*, *Molnár & Tóth*, 2009, 2010; *Csapó*, *Molnár*, *Pap-Szigeti & Tóth*, 2009; *Higgins*, *Russell & Hoffmann*, 2005; *Jeong*, 2014; *Tóth & Hódi*, 2011).

The inconsistency of the research results stem from the fact that the cited studies tested students in different assessment domains with various types of instruments with distinct softwares and software installations. They often focused on various samples of students and applied various data collection types and data analysis methods.

GOALS AND RESEARCH QUESTIONS OF THE EMPIRCAL RESEARCH

The objective of the empirical research presented in the dissertation is to examine the prevalence of mode effect in three assessment domains (inductive reasoning, reading comprehension and problem solving). This research attempts to provide a detailed examination on the effects of the characteristics of the assessment domain, sample and item. The dissertation does not aim at measuring the effect of infrastructural factors as the infrastructure available at Hungarian schools is constantly changing, renewing. Furthermore, through the development of online tests it was assured that every piece of information is presented in the same way both on the smallest and largest monitors available at schools.

When examining the characteristics of the assessment domains we aimed at comparing students' test achievements at test and subtest levels following international guidelines and recommendations (see e.g. *Bartram & Coyne*, 2005). Comparative analyses were carried out to examine the prevalence of the mode effect in each of the assessment domains. We compared the reliability of the paper-and-pencil and computer-based tests in order to show that students' performance may be measured with the same reliability in both test environments (*Lottridge*, *Nicewander*, *Schulz & Mitzel*, 2008).

This work presents comparative analyses concerning the sample characteristics drawing on students' background variables (e.g. school-grade, gender, place of living, socioeconomic status) and the variables related to computer use (computer familiarity, computer attitude and computer skills). We purpose to examine the mode effect assessments with the variables related to paper-and-pencil test results in which we supposed that the students with different abilities will react differently to the modification of the testing environment.

The empirical research was motivated by the need to take into consideration not only the mean test performance of students at test and subtest level, but also at item level. The reasons for this are as follows. (a) If there is evidence for mode effect of whole test or subtest level, the cause of mode effect can be revealed by the item level assessments; (b) Even in case of score equivalence, certain items may behave differently in the two environments (see *Gu*, *Drake & Wolfe*, 2006; *Poggio* et al., 2005; *Sandene*, *Horkay*, *Bennett*, *Allen*, *Braswell*, *Kaplan & Oranje*, 2005).

THE METHODS OF THE EMPIRICAL RESEARCH

Instruments

For this research we selected tests from the tests of the Hungarian Educational Longitudinal Program (*Csapó*, 2007), which were earlier validated and used on a large sample and could be digitalized without or with minimal modification. For the purposes of the mode effect study the validated inductive reasoning test developed by *Csapó* (1994) was selected in the first data collection. In the second empirical study, the reading comprehension test developed by *Éva Molnár* and *Gyöngyvér Molnár's* (2010) problem solving test were involved in the analyses. Besides the cognitive tests a background questionnaire was devised.

The inductive reasoning test

The inductive reasoning test (58 items) consisted of three subtests: number analogies (14 items), word analogies (14 items) and numerical series (16 items) (*Csapó*, 1994). The subtest of number analogies and numerical series contains short answer items whereas the subtest of word analogies contains simple choice items.

Reading comprehension test

Two parallel test versions of the reading comprehension measurement were applied. The two test versions contained the same instructions and the same item types. Both tests comprised 36 simple choice items (see *Tóth & Hódi*, 2011), had the same structure and both tests consisted of two subtests. The first subtest was a continuous text; the second one contained various non-continuous document types (tables, maps and diagrams).

The problem solving test

The problem solving test (28 items) presented a realistic situation in which the problems related to the vacation of a five-member family had to be solved by the students taking the test. The test contained simple choice and short answer items that required counting for the solution.

Background questionnaire

In order to map the characteristics of students we developed a questionnaire, in which we gathered information about background (e.g. age, gender, socio-economic status, number of books) and computer related variables. The questionnaire was grouped into two main components: background data (9 items) and factors related to the use of computers (56 items): computer familiarity, computer attitude and computer skills.

The computer experience sub-questionnaire (based on *Jones & Clarke*, 1995 definition) intended to reveal how long students have been using a computer, the frequency of computer use, how many hours students spend using a computer daily and for what purposes.

In order to assess students' attitude towards computers we adapted the first part of the Computer Attitude Questionnaire (*Knezek*, *Christensen*, *Miyashita* & *Ropp,* 2000), as part of which we mapped computer importance, computer enjoyment and computer anxiety with 18 items.

The computer skill sub-questionnaire contained 17 items. In this part we focused on the acquisition levels of the various computer operations.

Sample

The sample of the empirical studies comprised 5^{th}-7^{th} graders. Its composition and the method of sample selection are summed up in Table 1. Research was conducted by paper-and-pencil instruments of the MTA-SZTE Research Group on the Development of Competencies in the framework of the Hungarian Educational Longitudinal Program.

Table 1. Main characteristics of the studies presented in the dissertation

Assessment domain	*Grade*	*Date of data collection*		*Design*	*N*
		PP	*CB*		
Inductive reasoning	5	2008	2008	Repeated	843
Inductive reasoning	5	2008	2010	Matched pair	702
Inductive reasoning	6	2005	2010	Matched pair	262
Inductive reasoning	7	2010	2014	Matched pair	410
Reading comprehension	6	2009	2009	Repeated	449
Problem solving	6	2010	2010	Matched pair	342

Note: PP: paper-and-pencil; CB: computer-based test

The comparison of the online and paper-and-pencil test results was done through two different sampling methods: (a) the same student took the paper-and-pencil and online version of the same test (in the table we refer to it as repeated measure design); or (b) we compared the achievements in a computerized environment to results achieved earlier in the longitudinal measures with matched pair design.

Data collection

The paper-and-pencil data collection took part between 2008 and 2014. The computer-based data collection was carried out by means of the TAO (Testing Assisté par Ordinateur – computer-based testing; *Farcot* & *Latour*, 2008; *Tóth*, *Molnár*, *Latour* & *Csapó*, 2011) platform. Along with the online tests, the background questionnaire was also administered with the TAO CAPI platform (*Devosa* & *Vízvári*, 2011).

Methods for data analyses

The data analyses were carried out with classical and probabilistic test theory methods. Furthermore, we extended our comparative analysis with a data mining method. We used the SPSS data analysing software, the Conquest and R program package and the WEKA data mining software to analyse the data.

RESULTS

The results are presented in the three main dimensions of the empirical research: (1) the domain, (2) the sample and the (3) item characteristics.

Measuring mode effect in various domains

In accordance with *Lundy*'s (2008) study, the comparative analysis of the paper-and-pencil and computer-based test results in the three assessment areas did not identify notable reliability differences at test level (see *Molnár*, *Tóth* & *Csapó*, 2011). The results obtained from the subtest level analysis show that the verbal analogy subtest containing simple choice items had the same reliability in both testing environments. But significant difference was realized between paper-and-pencil and computer-based Cronbach-α values on subtests, which contained short answer items; the paper-and-pencil tests had higher reliability than the computer-based test (*Tóth*, 2014).

Based on the subtest level results achieved in reading comprehension, the non-continuous texts had the same reliability in both environments, whereas the continuous text had higher reliability in printed format than online (*Tóth* & *Hódi*, 2010). The problem solving test cannot be split into subtests, thus, we did not involve it in the subtest level comparison.

The analysis of paper- and computer-based performance differences justified small or medium mode effect in every measurement domain, and this finding implies that the test environment affects students' mean performance. Students' paper-and-pencil test scores were higher than the computer-based ones in every assessment domain (see *Csapó* et al., 2009; *Hódi* & *Tóth*, 2010; *Molnár* & *Tóth*, 2008; *Molnár*, *Tóth* & *Tóth*, 2010; *Tóth*, 2010, *Tóth*, *Molnár* & *Csapó*, 2009), which finding is consistent with findings of previous studies by *Choi* and *Tinkler* (2002) and *Pomplun*, *Ritchie* and *Cluster* (2006).

The subtest level analyses had similar findings to the comparative reliability analysis. The examination based on the subskills of inductive reasoning concluded that mode effect exists on subtests that contain short answer items, which require counting for the solution. Students achieved significantly higher scores on paper than online. But students' performance on the verbal analogy subtest containing simple choice items was test mode independent or only low effect was measured.

The analysis of the reading comprehension results showed that students' paper-and-pencil achievement was higher in an online environment in both text formats. Furthermore, a higher mode effect was found in the comprehension of the continuous text than in understanding its non-continuous counterpart (*Tóth* & *Hódi*, 2011).

The relationship between sample characteristics and mode effect

First we investigated the relationship between students' grade (age) and mode effect. Our conclusion is in accordance with *Kingston's* (2009) result. The longitudinal analysis in three school grades with matched pair design identified no change in mode effect neither at test level nor at subtest level.

We failed to find significant relationship between socioeconomic status and mode effect in the inductive reasoning domain. This finding is similar to the research conclusion reached by *Poggio* et al. (2005), *Sandene* et al. (2005), and *Horkay* et al. (2005). But in the problem solving domain we found a minor association between socioeconomic status and mode effect.

The mode. effect analysis carried out based on gender showed that the digitalization of paper-and-pencil instrument affects both boys' and girls' mean performance. Both boys and girls achieved higher test scores on paper than online. The mean performance differences of paper-and-pencil and computer-based assessment were equal at test and subtest level regardless gender. So, neither boys nor girls are advantaged or disadvantaged by the introduction of computer-based assessment.

The analysis of computer related variables revealed that the number of hours spent on computer a day showed a weak relationship with the mode effect at test and subtest level. However, those who had a computer at home – thereby had the possibility to use computer several times – were not advantaged in computer-based testing. We found no significant relationship between further computer experience variables (e.g. how long students have been using a computer and how often they do it and what they use it for) and mode effect, so students with various computer familiarity and skills reacted equally to the change of testing environment in the study.

Computer attitudes and the test mode effect variables were independent from each other, and the three subscales of computer attitudes (computer importance, computer enjoyment and computer anxiety) were also not related to the mode effect in the three assessment domains, which is consistent with the finding of *Lissitz*, *Jiao*, *Xie*, *Li* and *Kang* (2011). Unlike computer attitude, computer knowledge had a slight influence on the extent of mode effect in problem solving and inductive reasoning.

The results of the study among subgroups based on paper-and-pencil performance demonstrated that students with different ability levels react differently when the test medium changes (as experienced by *Pomplun* et al., 2006). Against our hypothesis and *Pomplun* et al.'s (2006) finding, students with lower ability level achieve higher performance in computerized environment than on paper on the same test in all three measurement domains (inductive reasoning, reading comprehension and problem solving). The achievement of mean ability level students was either test mode independent or the change of testing medium had only a small impact on test performance. Regarding the direction of the mode effect, students with average ability level performed slightly better in traditional paper-and-pencil environment than in an online environment. However, high ability level students are disadvantaged by digitalization of paper-based tests in all three measurement areas.

The results of item-level analyses

The aim of the item-level analyses was to examine the extent to which item type, item difficulty and item solution/completion activity is related to mode effect. The item-level data analysis revealed that the item difficulty indices of paper-and-pencil and computer-based tests typically differ significantly from each other. The examination of item types reflected that significant mode effect occurred on both (short answer and simple choice) item formats; solving both types of items was easier in paper-and-pencil environment than on computer. There is no notable difference between the extent of mode effects on short answer and simple choice items, so the direction and rate of mode effect is equal. The analysis of the relationship between item type and mode effect highlighted (in contrast with the findings of *Russell*, 1999 and *Russel* & *Haney,* 1997) that the typing of short answers associated with the familiarity of using keyboards does not produce mode effect.

There was a relationship between the item difficulty indexes of the paper-and-pencil tests and the extent of the item-level mode effect in all of the measured domains. Thus, item difficulty accounted for differential item functioning regardless of the domain.

The items were classified according to whether it was necessary to count in solving tasks. Then we examined whether the activity of solving tasks had a significant effect on the rate of mode effect. Based on the results, the items requiring no counting were test mode independent but the items which were solved by counting showed notable mode effect.

We examined the mean performance differences realized on problem solving and reading comprehension tests within text formats (e.g. table, figure). The items belonging to various text formats showed various extent of mode effect. The smallest mode effect was measured on items, which belong to non-continuous texts, more specifically, in comprehending information presented in tables. On the other hand, data gained from the study concerning problem solving showed the greatest mode effect in information retrieval tasks from tables. The difference of the diverse findings can be explained by the difference in test-taking activity. Students had to retrieve information in the reading comprehension test whereas in the problem solving domain students had to count with the information (numbers) retrieved from tables. These finding suggest that sixth graders are almost as capable of retrieving information online as on paper. However, the additional cognitive demand on the aspects (i.e. counting) may make answering the item more difficult and result in a greater mode effect. The difference in mode effect found in reading comprehension and problem solving imply that text format itself does not account for the mode effect.

In sum, the item type and the text format are not related with the rate of mode effect, but the item difficulty and the activity during test taking (requiring counting or not) had an impact on the extent of mode effect. But the question which variable has the highest information gain in predicting differential item functioning is yet to be answered. Therefore, in order to identify the attribute bearing the highest information gain in predicting differential item functioning decision tree method (C4.5 data mining algorithm) was applied. The variables (measurement area, sample design and item characteristics) involved in the study could predict the differential item functioning with 68% accuracy. Based on the output of the algorithm, the measurement area had the highest information gain. The item difficulty had also a prominent role in both problem solving and inductive reasoning. Furthermore, the

sample designs seemed to be an important factor in predicting differential item functioning as well.

FINDINGS AND FURTHER RESEARCH

The greatest importance of this empirical research is that we examined mode effect from three perspectives (the characteristics of the measurement area, sample and item) in order to be able to build a linear transformation between paper-and-pencil and computer-based test results. A key finding of the empirical research was that the data derived from the paper-and-pencil and computer-based assessment had the same reliability, so they can be applied to further educational research. The results of the thesis confirmed the finding of other international mode effect studies, which identified significant mode effect (at test, subtest and item level) in various measurement areas. Therefore, if we introduce online testing but we want to use the results previously collected in a paper-and-pencil form at the same time, it is desirable to have a linear transformation, which correspond the data/results of one test environment into the data of the other test environment in order to minimize the difference in performance that is caused by the change of test environments.

In sum, the studies introduced in the thesis aimed at carrying out analyses for a better understanding of the difference in students' performance when it comes to being tested in different environments in different domains. Therefore, we extended our analyses to criteria that are present and applicable in more or all domains. Our results may be generalizable and may provide a good starting point for further studies and analyses. However, a great limitation to our studies is that they do not provide an answer for the underlying theoretical issues of the certain domains. Thus, it is recommended to carry out further research in the following areas. Further experimental investigations are needed to study whether various types of problems have the same ratio of mode effect. Additionally, further research might explore the reason why students with various ability levels react differently when the test medium changes. It would also be worthwhile to extend the mode effect study in more age groups to map out how the introduction of computer-based testing influences students' achievement. These studies would advance the knowledge presently available in this domain.

References

Asuni, N. (2008). Quality Features of TCExam, an Open-Source Computer-Based Assessment Software. In Scheuermann, F., & Björnsson, J. (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 58–63). Luxemburg: Office for Official Publications of the European Communities.

Bartram, D., & Coyne, L. (2005). *International Guidelines on Computer-Based and Internet Delivered Testing*. Punta Gorda: International Test Comission.

Bennett, R.E. (2003). *Online assessment and the comparability of score meaning*. Princeton: Educational Testing Service. Retrieved from http://www.ets.org/

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode

effects in NAEP. *Journal of Technology, Learning, and Assessment*, *6*(9). Retrieved from http://www.jtla.org

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research*, *31*(1), 51–60.

Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting*. Retrieved from http://www.docstoc.com/docs/102661762/Evaluating-comparability-of-paper-and-pencil-and-computer--based

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, *33*(5), 593–602.

Csapó, B. (2007). Hosszmetszeti felmérések iskolai kontextusban – az első átfogó magyar iskolai longitudinális kutatási program elméleti és módszertani keretei. *Magyar Pedagógia*, *107*(4), 321–355.

Csapó, B. (1994). Az induktív gondolkodás fejlődése. *Magyar Pedagógia*, *94*(1-2), 53–80.

Csapó, B., Ainley, J., Bennett, R., Latour, T., & Law, N. (2012). Technological issues of computer-based assessment of 21st century skills. In Griffin, P., McGaw, B., & Care, E. (Eds.), *Assessment and teaching of 21st century skills* (pp. 143–230). New York: Springer.

Csapó, B., Molnár, Gy., Pap-Szigeti R., & Tóth K. (2009). A mérés-értékelés új tendenciái, a papír alapú teszteléstől az online tesztelésig. In Perjés István & Kozma Tamás (Eds.), *Új kutatások a neveléstudományokban. Hatékony tudomány, pedagógiai kultúra, sikeres iskola* (pp. 99–108). Budapest: Magyar Tudományos Akadémia.

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

Csapó, B., Molnár, Gy., & Tóth, K. (2008). A papíralapú teszteléstől a számítógépes adaptív tesztelésig. A pedagógiai mérés-értékelés technikájának fejlődési tendenciái. *Iskolakultúra*, Vol. 3-4., 3–16.

Devosa, I., & Vízvári G. (2011, April). *Recent Developments of TAO CAPI*. Paper presented at 3rd Szeged Workshop on Educational Evaluation. Szeged, Hungary.

Farcot, M., & Latour, T. (2008). An open source and large-scale computer based assessment platform: A real winner. In Scheuermann, F., & Pereira, A. G. (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.

Gaskill, J., & Marshall, M. (2007). *Comparisons between paper and computer-based tests: Foundation Skills Assessment 2001–2006 data*. Society for the Advancement of Excellence in Education, Kelowna. Retrieved from http://www.saee.ca/pdfs/038.pdf.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626.

Gu , L., Drake, S., & Wolfe, E. W. (2006). Differential item functioning of GRE mathematics items across computerized and paper-and-pencil testing media. *Journal of Technology,*

Learning, and Assessment, *5*(4). Retrieved from Journal of Technology, Learning, and Assessment [on-line] http://www.jtla.org

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning, and Assessment*, *3*(4). Retrieved from http://www.jtla.org

Horne, J. (2007). Gender differences in computerised and conventional educational tests. *Journal of Computer Assisted Learning*, *23*(1), 47–55.

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 from http://www.jtla.org

Hódi, Á., & Tóth, K. (2010, July). *Assessing Reading Skills in Printed and Computerized Environment*. Paper presented at Junior Researchers of the European Association for Research on Learning and Instruction (JURE) Conference. Frankfurt am Main, Germany.

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour and Information Technology*, *33*(4), 410–422.

Jones, T., & Clarke, V. A (1995). Diversity as a determinant of attitudes: a possible explanation of the apparent advantage of single-sex settings. *Journal of Educational Computing Research*, *12*(1), 51–64.

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). Retrieved from http://www.jtla.org

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 Populations: A Synthesis. *Applied Measurement in Education*, *22*(1), 22–37.

Knezek, G. A., Christensen, R. W., Miyashita, K. T., & Ropp, M. M. (2000). *Instruments for Assessing Educator Progress in Technology Integration*. Denton, Texas, USA: Institute for the Integration of Technology into Teaching and Learning University of North Texas. Retrieved from http://www.iittl.unt.edu/pt3II/book1.htm

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, *6*(1), 1–24.

Lissitz, R., W., Jiao, H., Xie, C., Li, M., & Kang, Y. J. (2011). *Student Characteristics and CBT Performance: An Overview of the Literature by MARCES*. Retrieved from MARCES, http://marces.org/current/Student%20Characteristics%20and%20CBT%20Performance.pdf

Lottridge, S. M, Nicewander, W. A., Schulz, E. M., & Mitzel, H. C. (2008). *Comparability of Paper-based and Computer-based Tests: A Review of the Methodology*. Pacific Metrics Corporation, Monterey. Retrieved from Pacific Metric, Corporation, http://www.pacificmetrics.com/research/white-papers/

Lundy, J. J. (2008). *Assessing psychometric equivalence of paper-and-pencil and interactive voice response (IVR) modes of administration for the EQ-5D and the QLQ-C30*. Retrieved from http://arizona.openrepository.com/arizona/bitstream/10150/193902/1/azu_etd_10040_sip1_m.pdf

Molnár, Gy. (2010). Papír- és számítógép alapú tesztelés összehasonlító vizsgálata problémamegoldó környezetben. In Perjés István & Kozma Tamás. *Új Kutatások a Neveléstudományokban* (pp. 135–144). Budapest: Aula Kiadó, Corvinus Egyetem.

Molnár, Gy., Tóth, K., & Tóth, E. (2010, October). *Developing Online Diagnostic Assessment - Experiences of a Large-scale National Case Study in Public Education in Hungary*. Paper presented at EDEN Workshop, Budapest, Hungary.

Molnár, Gy., Tóth, K., & Csapó, B. (2011, August). *The relationship between item type, students' characteristics and media-effect in CBA*. Paper presented at 14th European Conference for the Research on Learning and Instruction. Exeter, United Kingdom.

Paek, P. (2005). *Recent trends in comparability studies.* (Pearson Educational Measurement Research Report 05-05). Upper Sadle River: Pearson Educational Measurement.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment*, *3*(6) Retrieved from http://www.jtla.org

Pomplun, M., Ritchie, T., & Custer, M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment*, *11*(2), 127–143.

Puhan, P., Boughton, K., & Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing. *Journal of Technology, Learning, and Assessment*, *6*(3). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/issue/view/172

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*, *7*(20). Retrieved from, http://epaa.asu.edu/epaa/v7n20.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, *5*(3), Paper 186.

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*. (NCES 2005-457). Washington: U.S. Department of Education, National Center for Education Statistics.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "Intermediate Constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning, and Assessment*, *4*(6), 3–44.

Schröders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement*, *71*(5), 849–869.

Tóth, K. (2014, May). *Papír-ceruza és számítógépes induktív gondolkodás-tesztek reliabilitásának összehasonlító vizsgálata*. Paper presented at 12th Pedagógiai Értékelési Konferencia. Szeged, Hungary.

Tóth, K. (2009). Papír-ceruza és számítógépes tesztek eredményeinek összehasonlító vizsgálata. In Vajda Zoltán (Eds.), *Bölcsészműhely 2009* (pp. 125-136). JATEPress, Szeged.

Tóth, K., & Hódi, Á. (2011, April). *Comparing Students Reading Comprehension Achievement along Different Text-types in Paper-based and Computerized Environment*. Paper presented at American Educational Research Association - Annual Meeting. New Orleans, USA.

Tóth, K., & Hódi, Á. (2010). Olvasási képesség mérése számítógépes környezetben. In Perjés István & Kozma Tamás (Eds.), *Új Kutatások a Neveléstudományokban* (pp. 145–155), Budapest: Aula Kiadó, Corvinus Egyetem.

Tóth, K., Molnár Gyöngyér & Csapó, B. (2009, August). *Online Assessment of Reasoning Skills*. Paper presented at 13th Biennal European Conference for the Research on Learning and Instruction. Amsterdam, Netherland.

Tóth, K. Molnár, Gy., Hódi, Á., & Csapó, B. (2011, August). *Examining Media-Effect among Subgroups of Students with Different Ability Levels*. Paper presented at 14th European Conference for the Research on Learning and Instruction. Exeter, United Kingdom.

Tóth, K., Molnár, Gy., Latour, T., & Csapó, B. (2011). Az online tesztelés lehetőségei és a TAO platform alkalmazása. *Új Pedagógiai Szemle*, *61*(1–2–3–4–5), 8–22.

Tóth, K., Molnár, Gy., Sascha Wüstenberg, Samuel Greiff & Csapó, B. (2011, August). *Measuring adults' dynamic problem solving competency*. Paper presented at 14th European Conference for the Research on Learning and Instruction. Exeter, United Kingdom.

Tóth, K., Rölke, H., & Goldhammer, F. (2012a, July). *Investigating Test-taking behaviour in simulation-based assessment – Visual data exploration*. Paper presented at 4th International Conference on Education and New Learning Technologies, Barcelona, Spain.

Tóth, K., Rölke, H., & Goldhammer, F. (2012b, April). *Educational Process Mining-Clustering Students' Test-taking Behaviour in Internet-based Simulations*. Paper presented at 10th Pedagógiai Értékelési Konferencia, Szeged, Hungary.

Tóth, K., Rölke, H., Greiff, S., & Wüstenberg, S. (2014). Discovering Students' Complex Problem Solving Strategies in Educational Assessment. In Stamper, J., Pardos, Z., Mavrikis, M., & McLaren, B. M. (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 225–228)

Tóth, K., Rölke, H., Naumann, J., & Goldhammer, F. (2012, September). *Analyse des Problemlöseverhaltens in simulierten Hypertext-Umgebungen*. Paper presented at 48. Kongress der Deutschen Gesellschaft für Psychologie (DGPs), Bielefeld, Germany.

Tóth, K., Wüstenberg, S., Rölke, H., & Greiff, S. (2012, July). *Prediction of students' performance on test taking processes in Complex Problem Solving*. Paper presented at 30th International Congress of Psychology. Cape Town, South Africa.

Wang, S., Jiao, H., Young, M., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paperand-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, *68*(1), 5–24.

Wang, S., Jiao, H., Young, M., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, *67*(2), 219–238.

Publications in the topic of the dissertation created by the author

Csapó, B., Molnár, Gy., Pap-Szigeti R., & Tóth, K. (2009). A mérés-értékelés új tendenciái, a papír alapú teszteléstől az online tesztelésig. In Perjés István & Kozma Tamás (Eds.), *Új kutatások a neveléstudományokban. Hatékony tudomány, pedagógiai kultúra, sikeres iskola* (pp. 99–108). Budapest: Magyar Tudományos Akadémia.

Csapó, B., Molnár, Gy., & Tóth, K. (2010, April). *Implementing an Online Formative Assessment System: From Paper-Based to Computer-Based Testing*. Paper presented at American Educational Research Association - Annual Meeting, Denver, USA.

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

Csapó, B., Molnár, Gy., & Tóth, K. (2008). A papíralapú teszteléstől a számítógépes adaptív tesztelésig. A pedagógiai mérés-értékelés technikájának fejlődési tendenciái. *Iskolakultúra*, Vol. 3-4., 3–16.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626.

Hódi, Á., & Tóth, K. (2010, July). *Assessing Reading Skills in Printed and Computerized Environment*. Paper presented at Junior Researchers of the European Association for Research on Learning and Instruction (JURE) Conference. Frankfurt am Main, Germany.

Hódi, Á., & Tóth, K. (2009, November). *Olvasási képesség mérése számítógépes környezetben*. Paper presented at 11th Országos Neveléstudományi Konferencia, Veszprém, Hungary.

Molnár, Gy., & Tóth, K. (2008, November). *A számítógép- és papír-alapú tesztelés eredményeinek összehasonlító vizsgálata 5. évfolyamon*. Paper presented at 8th Országos Neveléstudományi Konferencia, Budapest, Hungary.

Molnár, Gy., Tóth, K., & Csapó, B. (2011, April). *Comparing Paper-Based and Computer-Based Testing in the First Grade*. Paper presented at American Educational Research Association - Annual Meeting, New Orleans, Louisiana, USA.

Molnár, Gy., Tóth, K., & Csapó, B. (2011, August). *The relationship between item type, students' characteristics and media-effect in CBA*. Paper presented at 14th European Conference for the Research on Learning and Instruction. Exeter, United Kingdom.

Molnár, Gy., Tóth, K., & Tóth, E. (2010, October). *Developing Online Diagnostic Assessment - Experiences of a Large-scale National Case Study in Public Education in Hungary*. Paper presented at EDEN Workshop, Budapest, Hungary.

Tóth, K. (2014, May). *Papír-ceruza és számítógépes induktív gondolkodás-tesztek reliabilitásának összehasonlító vizsgálata*. Paper presented at 12th Pedagógiai Értékelési Konferencia. Szeged, Hungary.

Tóth, K. (2010a, April). *Adatbányászat a neveléstudomány területén (Educational Data Mining)*. Paper presented at 8th Pedagógiai Értékelési Konferencia. Szeged, Hungary.

Tóth, K. (2010b, April). *Comparing Paper- and Computer-based Test-Taking Processes*. Paper presented at 2nd Szeged Workshop on Educational Evaluation. Szeged, Hungary.

Tóth, K. (2009). Papír-ceruza és számítógépes tesztek eredményeinek összehasonlító vizsgálata. In Vajda Zoltán (Eds.), *Bölcsészműhely 2009* (pp. 125-136). Szeged: JATEPress.

Tóth, K. (2009a, April). *The Szeged Experience*. Paper presented at Szeged Workshop on Educational Evaluation. Szeged, Hungary.

Tóth, K. (2009b, November). *Számítógépes és papír-ceruza tesztek eredményeinek összehasonlító vizsgálata háttérváltozók alapján*. Paper presented at 9th Országos Neveléstudományi Konferencia, Veszprém, Hungary.

Tóth, K. (2009c, April). *Az elektronikus tesztelés és tesztvégrehajtás fajtái*. Paper presented at 7th Pedagógiai Értékelési Konferencia, Szeged, Hungary.

Tóth, K. (2008, November). *Az online teszteléssel kapcsolatos attitűdök és eredmények a háttérváltozók tükrében*. Paper presented at 8th Országos Neveléstudományi Konferencia, Budapest, Hungary.

Tóth, K., & Hódi, Á. (2013). *A mérőeszköz-bővítéstől a tesztelési folyamat vizsgálatáig: számítógépes tesztelés nagymintás nemzetközi vizsgálatokban*. *Iskolakultúra*, Vol. 9, 75–88.

Tóth, K., & Hódi, Á. (2011, April). *Comparing Students Reading Comprehension Achievement along Different Text-types in Paper-based and Computerized Environment*. Paper presented at American Educational Research Association - Annual Meeting. New Orleans, Louisiana, USA.

Tóth, K., & Hódi, Á. (2010). Olvasási képesség mérése számítógépes környezetben. In Perjés István & Kozma Tamás (Eds.), *Új Kutatások a Neveléstudományokban* (pp. 145–155), Budapest: Aula Kiadó, Corvinus Egyetem.

Tóth, K., & Hódi, Á. (2010b, April). *Olvasási képesség fejlettségének mérése online környezetben kisiskolás tanulók körében*. Paper presented at 8th Pedagógiai Értékelési Konferencia. Szeged, Hungary.

Tóth, K., Molnár Gyöngyér & Csapó, B. (2009, August). *Online Assessment of Reasoning Skills*. Paper presented at 13th Biennal European Conference for the Research on Learning and Instruction. Amsterdam, Netherland.

Tóth, K., Molnár, Gy., & Csapó, B. (2008, April). *A számítógépes tesztelés lehetőségei*. Paper presented at 6th Pedagógiai Értékelési Konferencia, Szeged, Hungary.

Tóth, K. Molnár, Gy., Hódi, Á., & Csapó, B. (2011, August). *Examining Media-Effect among Subgroups of Students with Different Ability Levels*. Paper presented at 14th European Conference for the Research on Learning and Instruction. Exeter, United Kingdom.

Tóth, K., Molnár, Gy., Latour, T. & Csapó, B. (2011). Az online tesztelés lehetőségei és a TAO platform alkalmazása. *Új Pedagógiai Szemle*, *61*(1–2–3–4–5), 8–22.

Tóth, K., Molnár, Gy., Wüstenberg, S., Greiff, S. & Csapó, B. (2011, August). *Measuring adults' dynamic problem solving competency*. Paper presented at 14th European Conference for the Research on Learning and Instruction. Exeter, United Kingdom.

Tóth, K., Rölke, H., & Goldhammer, F. (2012a, July). *Investigating Test-taking behaviour in simulation-based assessment – Visual data exploration*. Paper presented at 4th International Conference on Education and New Learning Technologies, Barcelona, Spain.

Tóth, K., Rölke, H., & Goldhammer, F. (2012b, April). *Educational Process Mining-Clustering Students' Test-taking Behaviour in Internet-based Simulations*. Paper presented at 10th Pedagógiai Értékelési Konferencia, Szeged, Hungary.

Tóth, K., Rölke, H., Greiff, S., & Wüstenberg, S. (2014). Discovering Students' Complex Problem Solving Strategies in Educational Assessment. In Stamper, J., Pardos, Z.,

Mavrikis, M., & McLaren, B. M. (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 225–228)

Tóth, K., Rölke, H., Naumann, J., & Goldhammer, F. (2012, September). *Analyse des Problemlöseverhaltens in simulierten Hypertext-Umgebungen*. Paper presented at 48. Kongress der Deutschen Gesellschaft für Psychologie (DGPs), Bielefeld, Germany.

Tóth, K., Wüstenberg, S., Rölke, H., & Greiff, S. (2012, July). *Prediction of students' performance on test taking processes in Complex Problem Solving*. Paper presented at 30th International Congress of Psychology. Cape Town, South Africa.