UNIVERSITY OF SZEGED
FACULTY OF ARTS
DORCTORAL SCHOOL OF EDUCATION
INFORMATION AND COMMUNICATION TECHNOLOGIES IN EDUCATION

LÁSZLÓ HÜLBER

THE TRANSITION TO TECHNOLOGY-BASED ASSESSMENT

COMPARATIVE ANALYSIS OF PAPER- AND COMPUTER-BASED ASSESSMENT IN MATHEMATICS FROM GRADE 1 TO 6

THESIS OF THE DISSERTATION

SUPERVISOR:
GYÖNGYVÉR MOLNÁR

Szeged
2015

Topic and structure of the disseratation

Since the millenium there has been an increasingly emphathic demand for international and national assessments. The key elements of system adminsitration and decision-making based on facts and those of the tracking down of different interventions are provided by the indicators and measurement numbers originating from assessments. The staek-holders need information-rich reports of reliable quality from the micro-level classroom processes to macr-level public education management. (*Csapó*, *Molnár*, *Pap-Szigeti* & *R. Tóth*, 2009).

The inforcommuncation technologies (ICT) play an increasingly vital role in the world of education; thus it can be regarded as a natural consequence that the integration of technology has begun in the field of assessment. With the introduction of electronic testing not only does the medium of the test change but such ooportunities open up through ICT technologies that can start decisive reforms in the field of assessment (*Lent*, 2009).

Paper-based assessment has reached its limits; the entire range of the exploitation of the potential given by the test medium has come to an end (*Scheuermann* & *Björnsson*, 2009). Technology has such opportunity for prosperity in several areas that unambiguously designate the future of assessment, the roads of further development. It is worth underlining two areas for innovation: (1) as a result of the application of multimedia, interactive, dynamic elements and the authentic usability of ICT technologies 21st century skills become measurable. (2) the automatized correction mechanisms, the immediate provision of the results, the application of adaptive test algorythms, the more motivating testing environment and the extension of the sphere of participants involved in testing, the more efficient, adaptable achievement of the evaluation process can be realized (*Csapó*, *Ainley*, *Bennett*, *Latour* & *Law*, 2012).

Generation tehories have differentiated among 4 stages as far as technology-based assessment was concerned (*Bunderson*, *Inouye* & *Olsen*, 1989; *Bennett*, 1998, 2008; *Redecker* & *Johannessen*, 2013). Experts link the first generation to the efficiency of the assessment-evaluation process but the task of the third and fourth generation testing is considered as the integration of the holistic, personalized assessment into the learning process. The different generation theories agree that the transition into the era of embedded assessment demands a paradigm change in the field of assessment beside the development of technology. The milestone reform changes can be expected from this shift which would cause the parctice of learning and assessment to be interlinked (*Redecker* & *Johannessen*, 2013).

In spite of definitve future plans, the realization is a long process full of challenges whose first step is the transition from paper-based testing. Assessment institutes do not consider the migration to computer-based assessment an immediate final goal; therefore they approach the realization with transitional stages (*Lent*, 2009). Experts recommend the digitalization of paper-based test as the first step of the shift of traditional testing to computer-based testing (*Pommerich*, 2004). Besides keeping the original format from multiple points of view the two testing areas are comparable. Such assessment arrangement makes the execution of comparatistic analyses possible that makes the effect of media effect identifiable related to the test results. (*Clariana* & *Wallace*, 2002).

As far as the main topic of the dissertation is concerned the shift from paper-based testing to computer-based testingi is in the focal point. For the investigation of migration, the

analysis of a large-sample analysis of paper and computer-based test of mathematics amongst 1st and 6th graders is presented.

Considering the current trends a great deal of doctoral research will be done in the field of technology-based assessment and evaluation in the near future. This study tries to give a theoretical frame and starting point in general for the topic of technology-based assessment.

The 1st chapter achieves this goal which fixes the terminological system of the topic, shows the different grouping methods of the types of electronic testing. The assessment-evaluation done with different info-communication technologies has several forms of realization depending upon which the state and level of the assessment process it is used at and upon what these electronic solutions are used for by what stake-holders. (*Molnár*, 2010). The give part helps the placement of online testing in this study with respect to different categories. The advantage of implementing technology into assessment is mentioned by numerous studies using different types of emphasis, typically along the comparison to paper-based medium. Stepping beyond this dialectics, the dissertation groups and structures the opportunities into a hierarchical model and the advantages coming from this by mapping what kind of connection and mutual effect the different part elements are in with each other (1st, 2nd chapter). The third part of the first chapter shows the development process of technology-based assessment, the generations following one another and their characteristics and the expectations related to the future. This part historically positions computer-based testing and shows what previous events and future perspectives can be phrased in relation to the specific assessment type indicated in the dissertation. Knowing the developmental tendencies it will be characterized in both international and national relation where the different countries and assessment institutes are at in the shift to computer-based assessment: what the extent is of the spread of electronic testing.

The 2nd chapter focuses on the research problematics of the dissertation first. It shows that the changing of the testing medium is not only a technological and logistic question but it makes us face several psychometric challenges to be solved for the reliable application of electronic testing. Its goal is to map all the relations of the media effect to validity and applied technology, the parameters of the sample, the objectivity and reliability. It characterizes the nature of media effect, it divides it into components and shows the relations to one another. It fixes the presentation of the arrangement of research strategy of assessing media effect applied in the dissertation. One can learn the extent of the efficiency of the transmitting medium by comparing the results of the same paper-based and computer-based test. The question of equivalence is the next step when it is justified what kind of cases can occur and what effect they have. The question arises as well as to whether it is possible to talk about equivalence. The next chapter (Chapter 2.2) tries to give an answer to the problems and questions by showing and synthesizing the analysis of the results of previous media effects. First the entity of the comparative investigations relevant from the point of view of the dissertation from the studies of the past 35 years. Following this I will present and analyze the results on the features of technology, on the test and item parameters and on the characteristics of the sample by specifically emphasizing and detailing the analyses related to Hungary and mathematics. The third part of the chapter presents and systemizes the strategies of media effect. It will describe the profile of the two dominant structuring modes (independent and aligning samples): what sub-types it has, what processes can be applied, what reliabilities they have and whether they are suitable for drawing consequences. The last sub-chapter (Chapter 2.4.) deals with the

presentation of international practices. It will be asserted that the treatability of the media effect can be increased by following such practices. For the solution of the situation *Waters & Pommerich* (2007) suggests greater regulations and standards which would assure that the assessment conditions for the participants in electronic assessment are valid, reliable and objective.

The 3rd chapter presents the goals, hypotheses, research strategy of the empirical research by following the structuring style of a traditional dissertation. I phrase four overarching, complex goals which I divide into 22 hypotheses all in all with general comparative suppositions along the item parameters and the characteristics related to the sample. I go into details about the background data of the participants in the assessment which I compare with the data provided by the Central Statistical Office. I present the circumstances of the data-gathering process: request letters, registration process, directives related to the assessment and assessment instructions. I found it justified to present the latter process because little information is given with relation to what kind of preparation processes belong to administering electronic testing. I present and visually explain the processes used in sample-fitting and anchoring and the application of the Rasch-modell. A part of the dissertation of huge relevance for future media effect investigations is the parameter system which is suitable for the identification of the influencing power of the testing medium. The variable system orders the entity of the examinable item parameters to certain steps of the process of task solving: information processing, task solving (*Csíkos & Csapó*, 2011) and task-solving activity.

The last chapter focuses on presenting the usable results along hypotheses in a descriptive manner in the first step. This is followed by the detailed interpretation of results, reflection with the research, the parallels and opposites presented in the 2nd chapter. Visually rich material attempts to interpret the conlusions of the given part. The next chapter is of relevant practical relevance since on the basis of the synthesis of the theory, facts experienced duriong the investigation and the results of the research it phrases such suggestions related to online maths assessment that are on the one hand part of further investigations, on the other hand they contribute to the goodness indicators of electronic testing.

The background problematics of the empirical investigation

Apart from the numerous opportunities for innovation creating technology-based assessment makes different psychometric questions arise and it demands solution of problems. Upon Transmitting the test, administering the answers and their automatized correction the question arises what kind of modifying effects the application of technology exerts on testing, the results of the tests. Do we measure the same on paper and computer? Does the validity of the test change? Do all the students react similarly to the changes of the testing medium? Are there students that are affected with disadvantage?

Question regarding validity change depending on whether the assessed construct comprises technological literacy, tool using knowledge. This may not arise when paper-based tests are digitalized since the traditional testing medium may not compirise the construct realted to the use of technology. As a consequence of this the question in the dissertation is whether the validity of the test is modified or is supplemented by factors related to such technologies that are not part of tests assessing knowledge.

The origins of the questions related to the applied technologies are that computers available in schools are built up of various software and hardware elements. The questions presented here is whether the solutions transmitted on different configurations can be considered equivalent, and whether a group of test-takes suffer any disadvantage only because one or more parameters of the infrastructure are different. The investigation of the paramters separately is not sufficient since these elements are in mutual effect with each other and they influence together the variables determining the transmitting of the test. (e.g. appearance, speed).

In case such group of test takers exists that suffers disadvantages or enjoys advantages because of the introduction of technology then questions related to the fairness of the testing arise. From the point of view of impartiality and validity it is not imaginable that any body will get into disadvantage because of the change of the medium as far as neither teachers, institutions nor countries are concerned. With regard to both summative and diagnostic assessments it is not a wanted state that judgements will be made on the basis of invalid results (*Lent*, 2009).

Media effect investigations

The paper and computer-based comparative investigations are the first step to the learning of the new form of testing. In case it can be proved that the results scored on the two testing media are statistically equivalent or can be made equivalent then the applicability of electronic testing can be proved. (*Schroeders*, 2009). Az *International Test Commission American Educational Research Association* (AERA), the *American Psychological Association* (APA) and the *National Council on Measurement in Education* (NCME) stipulate at the guidelines related to the equivalence of the tests that the equivalence between the computer and paper-based fersions of test must be supported by documented evidence (*ITC*, 2006; *AERA*, *APA* & *NCME*, 1999, 2014). Based on the above documents and the studies conducted in the theme, the following points must be provided for the equivalence of a test taken on two different media (*AERA*, *APA* & *NCME*, 1999, 2014):

- the reliability of the computer and paper-based tests and the comparability of these reliability indicators (*ITC*, 2006);
- the equivalence of the validity of computer and paper-based testing (*Hargreaves*, *Shorrocks-Taylor*, *Swinnerton*, *Tait* & *Threlfall*, 2004);
- equivalence of item-level analysis justified in conjunction with tasks (*Puhan*, *Boughton* & *Kim*, 2007);
- the equivalanece provided by comparisons identifying the relevant background variables of the sample and individual differences (*Wolfe* & *Manalo*, 2005);
- the independence of the influencing effects of the applied technological parameters (*CTB/McGraw-Hill*, 2003).

In case equivalence can be stipulated based on the above points, then the two tests can be considered parallel and it is justified that neither modes influence the irrelevant variance of the construct. In case the equivalence cannot be proved then they do not assess the construct or the construct-irrelevant variance differs and both of them can be realized at the same time.

(*Csapó* et al., 2012). In case the new testing medium influences the results, it must be determined (*R. Tóth*, 2009):

- the extent of the difference between the results achieved in the two tesing media;
- the direction of the difference;
- the range of participants gaining advantage or suffering disadvantage;
- variables responsible for the difference

Research of media effect dates back to the same time as the history of technology-based assessment but no unitary conlusion in respect to all infrastructures, areas, contexts item formats can be drawn on the basis of the literature. In the background of the contradictions of the results lies the fact that different studies focused on different construct on a different sample with differing sample forming and they assessed with different itemformats on differing hardware/software infrastructure and they analyzed with diverse analyzing tecniques (*Wang & Shin*, 2009). The results of the investigations are very often questionable because of deficient documentation or because certain variables have not been given significance or have not been controlled for. Studies published before the 1990s can be considered out-of-date because differences are significant between the architectures used then and now or because present test-takes have tool-using experience above average. Because of he effect of the large amount of variable system and its differing result *Pommerich* (2004) suggests for the sake of responsible shifting that before shifting to electronic testing all assessment organizations should conduct comparative studies with their own test batteries, own technological background and sample because the results of the studies cannot be generalized entirely. It is typical of the nature of media effects that they are necessary all the time and the series of research studies do not come to an end since technology develops continuously (see touchscreen tools), new testing forms, items appear and the features of participants also change.

Research goals and hypotheses

The problem appearing in the dissertation not only sets a challenge at the level of theory but specific practical relevance can be matched to the set gooals. The goal of the Szeged research institute having 20 years of experience in the field of assessment (MTA-SZTE Képességkutató Csoport, SZTE Oktatáselméleti Kutatócsoport, SZTE Neveléstudományi Intézet és SZTE Neveléstudományi Doktori Iskola) is to make the data taking methods at research done in numerous fields technology-based. In order to conduct the migration responsibly, investigations comparing paper and compter-based assessment are necessary, which is the subject of the present dissertation. Apart from these specific conjunczions, the study might provide useful information for other assessment programs such the the media shift at the National Competence Assessments and for futre computer-based tests.

In the field of maths the dissertation compares the results taken paper and computer at 1^{st}-6^{th} grade. The firld of mathematics was chosen because it belongs amongst basic school subjects and it determines the students' school progress from the beginning of primary school to the end of secondary school. (*Bennett*, *Braswell*, *Oranje*, *Sandene*, *Kaplan & Yan*, 2008).

The interval between 1^{st} and 6^{th} grade covers the age spectrum when in the students' skillfulness, aptritude in computers huge differences can exist, thus the role of age and ICT

experience can be examined well.The goals of the research are of general phrasing: they overarch more or concerning the first all the hypotheses.

Goals of the research:

- mapping the 1st and 6th graders' behavior in an online testing environment
- comparing the achievements in paper and computer-based environment
- determining item parameters at which the differeing or the equivalence of the achievements of a certain direction is typical
- identifying such sub-samples which typicly behave identically or differently in different test environments

The hypothses of the research phrase conditions along parameters related to items and samples of general orientation.

Hypotheses of general orientation:

- Related to reliability:
    - H1: The interior consistency of online maths tests is appropriate. (Cronbach-$\alpha \geq 0,8$).
- Related to the average differences between paper and computer-based test results
    - H2a: at the level of average differences there are no significant differences between the results assessed on the two media.
    - H2b: the differences between the two media dependent on the sample and the item parameters are determined.

The hypotheses related to the charactersitics of the items:

- Related to the type of tasks:
    - H3a: In case of clos-ended item types, studentds achieve better at the computer..
    - H3b: In case of open-ended text-producing tasks, the paper-based method proves to be easier for students.
- Related to the internal features of tasks and the knowledge elements necessary for solution:
    - H4a: The content dimension of the tasks can influence the amount of differences.
    - H4b: Tasks demanding higher level thinking functions achieve better at paper-based tests.
- Related to the pieces of information pertaining to the tasks:
    - H5a: The amount of textual information (character number) appearing at tasks correlates with the amount of the media effect.
    - H5b: The items using graphic elements are more successful at computer-based tests.
    - H5c: The application of tables does not influence media effect.

    - H5d: The organization system of information does not influence the difference between the two testing media.

- Related to the features of replies required by the task:
    - H6a: The length of the reply required by the tasks correlates with success at the computer.
    - H6b: Media effect can be experienced at tasks requiring kexboard combinations.

Hypotheses related to the sample:

- Related to the role of age:
    - H7a: With the passing of age the difference between the results scored on paper and computer-based tests decrease.
    - H7b: Significant differences exist between creation grades in the extent of differences.

- Related to the sex of the students:
    - H8: The sex of the students influences the amount of the differences between achievements.

- Related to the geographical parameters of the students
    - H9: The residential location (region) of the students does not modify achievement with respect to certain media.

- Related to the education of the students' parents:
    - H10a: The education of the mother is a background variable influencing media effect.
    - H10b: The education of the father is a background variable influencing media effect. A tanulók tanulmányi előmenetelével kapcsolatban:

    - H11a: The mathematics scores of the students (semester mark) do not influence the differences between results reached in certain testing environment.
    - H11b: The mean of the students' school marks (semester mark) does not influence the differences between results in certain testing environment.
- Related to the ICT training of the students:
    - H12: Those students that have an opportunity for ICT classes and got to extra-curricualr informatics classes are more successful than their peers on the online mathematics test.

The research strategy of the investigation

Participants

The sample of the PP study was drawn from 1 to 6 grade primary school students, representative of the school population in Hungary (n=40 571). The computer-based version of the tests was administered two years later to the sub-sample of the original sample (n= 22 715), with at least 3000 students from each grade. The similarity of the samples of the two time points was ensured by matching participants on the student level, as explained below. Accordingly, the two independent, but matched samples can be considered identical samples.

The online testing was carried out via the eDia (Electronic Diagnostic Assess-ment; *Molnár*, 2013) platform through Internet. The data collection took place in the computer labs of the participating schools, using the available computers and browsers installed. Before the testing we drew the supervising teachers' attention to the possibility of giving the students papers to make notes if necessary.

Instruments of the study

On average we selected 30 paper-based test versions in each grade. Each version was made up of three clusters. One cluster consisted of 3-4 pieces of Mathematics tasks with 4-5 items, making about 15 items in total. With regard to the paper-based tests we prepared 10 versions that were as similar as possible to the computer-based version regarding the diversity and the even element number according to the task types and item parameters. In order for more fine-grained analyses, during the se-lection of tasks special emphasis was put on presenting a large number of graphic items (*Hülber*, 2012). During the digitalization of the tasks, their characteristic fea-tures and appearance was kept. On the whole 184 tasks (879 items) were presented parallel on both medias.

Apart from the Mathematics tests, every student filled in a questionnaire consisting of 7 items regarding background information (gender, qualification of parents, school progress, progress in Maths; Maths attitude, participation in IT classes in school).

Design

We applied a between-subject design. As the characteristic features of the sample of the two data collections were different, getting reliable results we adjusted the sample in the paper-based testing in order to match the sample of the computer-based assessment. To every student participating in the computer-based data collections, we matched a student from the paper-based sample with similar characteristics regarding grade, gender, qualification of parents and grade point average. As a result of this adjustment at the student level, we could match pairs satisfying at least 5 aspects in 70% of the students.

The comparison of the performance shown on the computer- and the paper-based tests was made possible by the anchor items and by applying latent-trait theory models. The special arrangement of the clusters of items ensured the stability of anchorage. All the 10 clusters used on computer were included in three different test versions, once in the beginning of the test, once in the middle and once in the end. Thus, all the 10 tests were anchored with all the other tests and the possible effect of the task's position was excluded in the test. The anchorage of the paper-based test happened with a similar technique.

The comparison of performance was carried out on different tests that were linked with anchor items and we implemented their conversion to a common scale and the ranking of the data with the help of the two-dimensional Rasch-model, supposing that despite the similarities of the tasks it is not necessarily true that a completely identical construction is measured on paper and on computer. To describe the reliability of tests we used the person-separation reliability index in addition to the analogous Cronbach's alpha. The person-separation reliability index is suitable for giving a common description of tests linked with anchor items.

To describe the behavior of the iems I created an own universally applicable perspective system in which I ordered the range of examinable itemparameters to single steps of the complex process of task solving. The model comprises the following elements: (1) the processing of the information pertaining to the tasks (information type, quantity, its organization), (2) the psychical, content, functional contextual features plaing a role during task solving (*Csíkos & Csapó*, 2011; *Vidákovich*, 2012), and (3) itemparamters related to the task solving activity, the task type, the amount of information wished to be fixed, its type and method of tool use.

Results of the research and their interpretation

(H1) The interior consistence of the online tests were acceptable in every grade and in case of every test version. The Cronbach-α indices had a higher value than 0.8. The means created from the indices are around 0.9, the person separation (EAP\PV) reliability indices determined by the Rasch-model also proved to be reliable and they took a higher value than 0.85 for all grades.

No differences, patterns can be observed between the indices in terms of grade; in case of even the smallest the values are excellent. This gives the statement the basic pillar that online testing can reliably be applied from the beginning of school from 6-7 years of age.

(H2) No significant difference appeared at the levele of means between the difficulty values of the itmes but these results in themselves are not proper to create a final judgement in conjunction with the comparison between paper and computer-based testing. (*Wilhelm & Schröeders*, 2008). Groups of items and samples exist where significant dofferences were found.

(H3a) In case of the first three grades within the close-ended tests significant difference was found for the computer-based medium at the tasks requiring alternative choice. Az első három évfolyam esetén a zártvégű feladatokon belül az alternatív választás típusú ($|t_1|=4.16$, $p<.01$; $|t_2|=2.67$, $p<.05$; $|t_3|=4.16$; $p<.01$). In the background of the data lies tha fact that an average 155 more relies were recorded on computer in electronic testing environment. This can be paralleled to the research results of *Mazzeo & Harvey* (1988), *Csapó* et al. (2009), *Johnson & Green* (2006), who found the ratio of replying attemps more frequent. It could be an explanation that students find compute-based testing less threatening, therefore they dare to try more bravely, more replies are given which altogether result in more test points. The consequence of this strategy dependent on testing environment is the guideline appearing at paper-based tests according to which it is an advisable state to apply various types of tasks.

(H4) In conjunction with the content elements and the psychic functions necessary for task solving, tasks containing statistical, combinatorical, higher-level functions prove to be more difficult on a computer basis than on paper basis. At the other subtypes no difference was

found with respect to certain media. Following the explanations of *Johnson & Green* (2006), *Hülber* (2012) and *R. Tóth & Hódi* (2011) these types of tasks require the use of notepapers to the greatest extent, the outlining of the problem, visual thinking experssable with a certain type of handwriting which cannot be achieved on computer. The problem is probably the fact students do not have enough experience in computer-based testing, do not have solving strategies and routine. As a result of this, they try to do more function by heart and would not like to adapt the computer screen to the note paper, copy data and divide their attention. The assumption is confirmed by the fact that supervising students suppose the use of note papers was very rare.

(H5) The amount of textual information (character number) appearing at the tasks influenced the results the same way both on computer-basis and paper-basis. The organizing method of the information and the application of tables did not influence the differences between the media. As regards the type of information, at tasks containing graphical elements a significant difference was found at 1st grade for the electronic medium. ($|t|=2.01$, $p<.05$). The studies of *Richardson*, *Baird*, *Ridgway*, *Ripley*, *Shorrocks-Taylor & Swan* (2002), and *Hülber* (2012) stipulated results similar to this. Causes lying in the background are the nicer appearance on computer, more life-like design which supposedly elicit greater attention and motivation (*Gyarmathy*, 2012) and children take more time to process them; thus they solve them better by interpreting them more deeply. Probably, in not every csa were the tests printed in color on paper whereas in case of computer-based assessment all the monitors were in color.

(H3b, H6) At the level of means no difference was found between the achievements on the open-ended tasks and the paper-based medium. The length of the answer given to the tasks also identically correlated with the item difficulty pertaining to the tasks in the two testing environment. It could be supposed on this bases that the application of the tasks meant a same amount of challenge in both media; however we do not know whether the typing and the writing of the tasks took the same amount of time. It could be supposed based on the feedback given by the supervising teachers that in case students know the answer, they struggle with the typing of the correct answer; thus no difference is shown at item level but on account of time loss they might suffer disadvantage. Tasks demanding keyboard combinations appearing from 3rd grade those working on the computer suffered a significant disadvantage.

(H7) From the perspective of foretelling differences one of the most determining variable is the age of the student. While in the first three years correlation values are around 0.7 between the item difficulty values appearing in the two media ($r_1=.7$; $r_2=.72$; $r_3=.69$), from 4th grade on an increasing correlation of the results can be observed ($r_4=.77$; $r_5=.85$). This reaches the 0.92 value which supposes a strong relationship between the two media. Among the correlation coefficients of 1st-4th, 4th-5th, 5th-6th grade the z-probes gave evidence of significant differences. The influencing strength of age or the significant differences between differences at the youngest age-group are justified by several studies (eg. *Choi & Tinkler*, 2002; *Ito & Sykes*, 2004; *Applegate*, 1993; *Barnes*, 2010). These studies unanimously attribute the lack of literacy in technology to the fact that the youngest students achieve worse on an electronic basis. At the age of 6-7 it is common that students have not used a computer or the use of a mouse is not evident for them to click in a textfield in order to write, what the role of ‘enter’ and ‘space’ is, how theys can get to the next test question or what the method is of typing in special characters. Their lack of experience in practice is justified by the fact that movement in the text amongst

1st and 2nd graders is rare, they do not return to pervious tasks, do not check, modify, think over their solutions but they go through them once which is probably in contrast to paper-based strategies and less successful. By every schoolyear the frequency of moving in a text increases by 10-15% which probably gets closer and closer to the paper-based mechanisms. Suppositions related to lack of experience in practice, unskilledness in technology are confirmed by the feedbacks of test supervisors.

(H8) With respect to sex, difference was found only at 1st grade. Regarding sex, difference was found only at 1st grade: girls achieved significantly worse on computer than on paper ($|t|=2.11$, $p<.05$). Similar results were found by *Horne* (2007), *Johnson* & *Green* (2006), *Halldórsson*, *McKelvie* & *Björnsson* (2009), *Csapó* et al. (2009). The following explanations can be made: boys achieved better because they are more motivated in a computer environment because they have bigger experience and as a result skillfulness and they are more confident. It can be further assumed that boys find it more important to acieve better on computer than on paper compared to girls who behave the other way around.

(H9) Mathematicly significant differences were found at only the first three grades in conjunction with the geographical location of the students (region) but no systematic organization of these differences can be observed. On the basis of these I do not find geographical location a predictive variable from the perspective of differences. On the basis of paper-based results we know that the differences in achievements pertaining to regions and counties generally belong to the effect of one variable which confirms the previous statement.

(H10) At the first three grades difference was found in conjuction with the influencing strength of the parents' education. Children of parents having college-level education significantly performed worse on the computer-based test. Similar results are found in the studies by *Bennett* et al. (2008) and *Hódi* and *R. Tóth* (2009). It can be phrased as a reason that students with better achievement react more sensitivel to the shift in the testing medium.They have bigger need to achieve well, the motivation for better achievement, thus it can give rise to bigger disturbance that they cannot use the habitual, successful task solving strategies and mechanisms.

(H11) From the perspective of learning achievements the age group sensitive to change could not be examined since students receive marks only at the end of the second grade. The drawing of conclusions was also exacerbated by the fact that few sample elements were at our disposal with respect to bad marks. From grade three to grade six in case of students with autumn semester mark 3 or hihger and of math mark 2 or higher no significant difference was found except for one case (4th grade semester mark average 3). Due to the listed deficiencies full investigation of the hypotheses could not be executed and the results refer to the fact that learning achievement would influence the difference between the media from grade three.

(H12) At the time of data-taking the students did not have informatics as a mandatory subject below 5th grade. In the first four schoolyears, huge samplesize differences are found with respect to whether students learn informatics or not. At 3rd and 4th grade having levelled sample size the fact the students has informatics or not at school had no effect.

Suggestions, futher research opportunities

The basis of the suggestions are the synthesis of the literature, the experiences of practical realizations, results, the feedbacks of supervisors participating in the assessment, system hosts and assessment coordinators. The practical relevance of the study is increased by these suggestions which might serve the more reliable, valid and objective realization of computer-based testing independent of the used testing platform and assessment area. My recommendations have been fixed around certain single problems.

- For the the handling of problems pertaining to the technological literacy of the students:
  - doing tutorials presenting the use of computers Számítógép használatát bemutató, gyakoroltató tutorialok (teaching programs)
  - doing sample and try-out tasks
- related to the prediction of media effect:
  - collecting contextual data and their analysis. For example how much tiem is needed to do a task related to the moving of a mouse, clicking with it and typing in texts.Tutorial, sample tasks and the whole test can be well suited to gather such information.
- For the handling of problems related to technology, infrastructure:
  - use of hardware, software diagnostic tools which by running in the background assess and take data of the features of the computer used for testing.
  - building up a database making possible the identification of computers. Apart from the identification of students and test results the identification of computers would be stored (through this providing the control of test results)
  - the process of certifying and classifying infrastructure through experts on the basis of international guidelines and instructions.
- for the task types demanding drawing related to their computer-based realization:
  - transforming of tasks, making them close-ended by outlining certain variations (it migh modify validity)
  - use of a digital pen
- for the aiding of students in shifting to online testing:
  - use of spectacular, ergonomical timing tools, for example: task grid
  - displaying time spent on test
- for the pasting special characters and character forming
  - application of tutorial
  - placing dedicated buttons, icons on the testing surface
- for the ergonomics of the testing surface:
  - creating a child-friendly test surface
  - feedback of results for the students in an interpretable, playful form.
- related to the break in the testing period:
  - providing the possibility of continuing the test immediately after reloading, restarting.

- working out such action plan which gives the action series to be followed for all abnormal cases

- for the aiding of operative functions preceding testing:
  - provodng a brochure colored with rich illustrations

- related to the structuring of the test:
  - application of task samples (space, width, interline spacing, font size, font tpye, align, sizes of textboxes, pictures etc.).
  - controlling mechanism, provodong pre-alarm warning in case of a display differing from this with respect to the media effect.

- for the negative effect originating from the élack of experience in practice in assessment on a compter base:
  - popularizing technology-based testing at a micro-level and in classroom practice as well.
  - providing cooperation with the teachers, holding further trainings, providing a platform suited for autonomous testing.

The majority of suggesstions contain possibilities of new research opportunities which make it possible to justify or refute the effect of the recommended processes on the goodness indices of testing. The study outlines further research directions related to the items with the open-ended tasks (more specifically to their task solving time) interpreted on the sample along the school achievements of the students and the processing of technological parameters is worthy of focusing on.

The significance of the dissertation and its limitations

The dissertation gives account of the experiences of the first country-level online assessment large sample size. The research can be considered unique and ground-breaking since it investigated primary school students of heterogenous ICT literacy on a large sample size on the basis of technology in the field of mathematical literacy that is a determining factors from a school achievement perspective. The analyzed items were not restricted to a narrow sphere (for example according to type: multiple-choice items) but they nrealy covered the entire parameter system characteristic of the items. The study used modern procedures: test anchoring, sample adaptation; the basis of mathematical analyses are characterized by calculations of probability assessment theory. The dissertation intends to provide a starting point on a theoretical basis for computer-based assessment with the presentation and synthesis of terminology, marking methods, possibilities and advantages, gerenation theories. Similar to media effect investigations it strives for the maximum by describing numerous studies from multiple perspectives highlighting the field of mathematics and Hungarian experiences. It also clarifies the methodological questions, types and characteristics. For the comparison of the items, such a complex model is presented which could mean a strating point of compratative investigations of different context.The practical relevance of the disseratation is increased by further suggeestions independent of platform.

The compratative investigation took place in the field of mathematics, thus the results can be generalized to this literacy field. Other areas might result in differing variable system not only related to content area. For example on the field of reading comprehension the processing of larger amount of information or the solution of tasks demanding longer answers might result in a media effect of different chacatereristic (*R. Tóth & Hódi*, 2011). The limitations of the generalizability of the items stem form the fact that there was no opportunity to observe the parameters of the applied computers during research; not an identical sample solved the tests on computer and on paper; not the same sample solved the tests on a paper basis and on a computer basis; different school grades are of different abilities compared to each other, regarding certain parameters they have a distribution of different sample size; due to the inequal distribution of the information bo calculation could be performed at student ability levels.

Literature

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999, 2014). *Standards for educational and psychological testing*. Washington, Amerikai Egyesült Államok.

Applegate, B. (1993). Construction of geometric analogy problems by young children in a computer-based test. *Journal of Educational Computing Research*, *9*(1), 61–77.

Barnes, S. K. (2010). Using computer-based testing with young children. Paper presented at the NERA Conference Proceedings 2010. Paper 22.

Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing*. Policy Information Center, Educational Testing Service, Princeton, NJ.

Bennett, R. E. (2008). *Technology for large-scale assessment*. ETS Report No. RM-08-10. Educational Testing Service, Princeton, NJ.

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), 4–38.

Bunderson, V. C., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In: Linn, R. L. (eds.): *Educational Measurement*. Macmillan, NewYork. 367–407.

Choi, S. W., & Tinkler, T. (2002). Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting. Előadás. In: Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Clariana, R., & Wallace, P. (2005). Test mode familiarity and performance-gender and race comparisons of test scores among computer-literate students in advanced information systems courses. *Journal of Information Systems Education*, *16*(2), 177–182.

CTB/McGraw-Hill (2003). *The computer-based or online administration of paper-pencil tests*. CTB/McGraw-Hill, Amerikai Egyesült Államok.

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

Csapó Benő, Molnár Gyöngyvér, Pap-Szigeti Róbert & R. Tóth Krisztina (2009). A mérés-értékelés új tendenciái: a papír és számítógép alapú tesztelés összehasonlító vizsgálatai általános iskolás, illetve főiskolás diákok körében. In: Perjés István és Kozma Tamás (szerk.): *Új kutatások a neveléstudományokban. Hatékony tudomány, pedagógiai kultúra, sikeres iskola*. Magyar Tudományos Akadémia, Budapest. 2009. 99–108.

Csíkos Csaba & Csapó Benő (2011). A diagnosztikus matematika mérések részletes tartalmi kereteinek kidolgozása: elméleti alapok és gyakorlati kérdések. In: Csapó Benő & Szendrei Mária (eds.): *Tartalmi keretek a matematika diagnosztikus értékeléséhez*. Nemzeti Tankönyvkiadó, Budapest. 59–99.

Gyarmathy Éva (2012). Ki van kulturális lemaradásban? In: Tóth-Mózer Szilvia, Lévai Dóra & Szekszárdi Júlia (szerk.): *Digitális nemzedék konferencia 2012 Tanulmánykötet*, ELTE Eötvös Kiadó, Budapest. 9–12.

Halldórsson, A., McKelvie, P., & Björnsson, J. (2009). Are Icelandic boys really better on computerized tests than conventional ones: Interaction between gender test modality and test performance. In: Scheuermann, F. & Björnsson, J. (eds.): *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*. Office for Official Publications of the European Communities, Luxemburg. 178–193.

Hargreaves, M., Shorrocks-Taylor, D., Swinnerton, B., Tait, K., & Threlfall, J. (2004). Computer or paper? That is the question: Does the medium in which assessment questions are presented affect children's performance in mathematics? *Educational Review*, *46*(1), 29–42.

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

Hülber László (2012). A papír és a számítógép alapú tesztelés összehasonlító vizsgálata különböző item paraméterek mentén. *Iskolakultúra*, *12*(12), 13–26.

Hülber László & Molnár Gyöngyvér (2013). Papír és számítógép alapú tesztelés nagymintás összehasonlító vizsgálata matematika területén, 1-6. évfolyamon. *Magyar Pedagógia*, *113*(4), 243–263.

International Test Comission (ITC) (2005). *International guidelines on computer-based and internet delivered testing*. International Test Commission. Bruxelles.

Ito, K., & Sykes, R. C. (2004). Comparability of scores from norm-reference paper-and-pencil and web-based linear tests for grades 4–12. Paper presented at Annual meeting of the American Educational Research Association, 12–16. 04. 2004., San Diego, USA.

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), 4–33.

Lent, v. G. (2009). Risks and benefits of CBT versus PBT in high-stakes testing. In: Scheuermann, F., & Björnsson, J. (eds.): *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*. Office for Official Publications of the European Communities, Luxemburg. 83–91.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature*. College Entrance Examination Board, New York.

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

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment*, *2*(6), 1–45.

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), 1–21.

Redecker, C., & Johannessen, Ø. (2013). Changing assessment – Towards a new assessment paradigm using ICT. *European Journal of Education*, *1*(48), 79–96.

Richardson, M. T., Baird, J-A., Ridgway, J., Ripley, M., Shorrocks-Taylor, D., & Swan, M. (2002): Challenging minds? Students' perceptions of computer-based World Class Tests of problem solving. *Computers in Human Behavior*, 18(*6*), 633–649.

R. Tóth Krisztina (2009). Papír-ceruza és számítógépes tesztek eredményeinek összehasonlító vizsgálata. In: Vajda Zoltán (eds.): *Bölcsészműhely 2009*. JATEPress, Szeged. 125–136.

R. Tóth Krisztina & Hódi Ágnes (2011). Számítógépes és papír-ceruza teszteredmények összehasonlító vizsgálata az olvasás-szövegértés területén. *Magyar Pedagógia*, *111*(4), 313–332.

Scheuermann, F., & Björnsson, J. (2009). *The transition to computerbased assessment: New approaches to skills assessment and implications for large-scale testing*. Office for Official Publications of the European Communities, Luxemburg.

Schroeders, U. (2009). Testing for equivalence of test data across media. In: Scheuermann, F., & Björnsson, J. (eds.): *The transition to computerbased assessment: New approaches to skills assessment and implications for large-scale testing*. Office for Official Publications of the European Communities, Luxemburg. 164−170.

Vidákovich Tibor (2012). A feladatok paraméterezése. Manuscript.

Wang, S., & Shin, C. D. (2009). Comparability of computerized adaptive and paper-pencil tests. *Test, Measurement & Research Service. Bulletin*, *13*. 1–7.

Waters, S. D., & Pommerich, M. (2007). Context effects in internet testing: A literature review. Paper presented at 22nd Annual Conference of the Society for Industrial and Organizational Psychology, 2007. április 7., New York City, Amerikai Egyesült Államok.

Wilhelm, O., & Schroeders, U. (2008). Computerized ability measurement: Some substantive dos and don'ts. In: Scheuermann, F., & Pereira, A.G. (eds.): *Towards a research agenda in computer-based assessment. Challenges and needs for European Educational Measurement*. European Commission Joint Research Centre, Ispra. 76-84.

Wolfe, E. W., & Manalo, J. R. (2005). *An investigation of the impact of composition medium on the quality of TOEFL writing scores*. Educational Testing Service, USA.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest: Generalised item response modelling software*. ACER press, Melbourne.