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

Andrea Magyar

Comparing the Measurement Effectiveness of Computer-Based Linear and Adaptive Tests

Dissertation Theses

Supervisor:
Gyöngyvér Molnár, PhD, habil., Associate Professor

Szeged
2015

Introduction

By the end of the 20th century, the most accepted and widespread paper-and-pencil tests (PP) had faced an increasing number of limits and the different opportunities of paper-based tests had gradually been exhausted (Molnár & Magyar, 2015). To proceed and to satisfy the needs in measurement and evaluation of the 21st century, a basic and qualitative change is required (Scheuermann & Pereira, 2008; Beller, 2013). The rapid development of technology clearly points the way forward to the transition to computer-based testing (Scheuermann & Björnsson, 2009; Molnár, 2010; Csapó, Ainley, Bennett, Latour & Law, 2012), which offers a wide range of new opportunities compared to paper-based testing, for example, a more motivating environment (Thompson & Pometric, 2007), the possibility of immediate evaluation (Wang, 2010), the display of innovative, multimedia elements with dynamically changing items (Greiff, Wüstenberg & Funke, 2012) and the implementation of personalized, adaptive testing methods (Eggen & Straetmans, 2000).

In the administration of computerized adaptive tests (CAT; Weiss, 2011), the order of the items is not predetermined in a fixed sequence; they are selected from an item pool such that the selected items are matched to the ability level of the individual based on his or her performance on items presented previously. For example, in the case of item-level adaptivity, if the student is able to complete the item correctly, a more difficult item is administered in the following step; if not, then he/she is provided with an easier one. By applying this algorithm to the whole test, a certain ability level can be assigned to each student at the end of the testing process. At that level, there is a high probability of his/her being able to complete the easier items correctly and the more difficult items incorrectly.

This type of administration and test assembly method facilitates much more precise performance measurement than the traditional linear approach, in which the items and the orders of the items are identical for everyone (Linacre, 2000; Magyar & Molnár, 2013). The available amount of information about items and individuals significantly increases (Miller, 2013; Magyar & Molnár, 2013). The probability that the individuals receive the same items in the same order becomes negligible, thus increasing test security (Wainer, 2000). All these properties create new opportunities in the field of measurement and evaluation. If we do not strive to extract extra information from the test, the number of items administered or the length of the test (Thompson & Way, 2007) decreases; in parallel, testing time is also reduced to a considerable extent—by half, on average (Frey & Seitz, 2009; Frey, Seitz & Kröhne, 2011).

We are currently in a transition phase; comparative studies between the traditional linear method and adaptive testing are therefore justified. They play a prominent role in longitudinal studies, where the results from earlier paper-based tests are compared to those from computer-based tests as well as comparing results when the two types of media testing are administered alternatively and simultaneously.

During the transition from traditional testing to adaptive testing comparative, a number of analyses of the two forms of tests have been conducted (Al-A'ali, 2007; Brossman & Guille, 2014; Crotts, Zenisky, Sireci & Li, 2013; Frey, Seitz & Kröhne, 2011; Guille, Becker, Zhu, Zhang, Song &Sun, 2011; Hambleton & Xing, 2006; Jodoin, Zenisky & Hambleton, 2006; Kingsbury & Hauser, 2004; Olea, Revuelta, Ximénez & Abad, 2000; Pyper & Lilley, 2010; Rotou, Patsula, Manfred & Rizavi, 2003; Thompson & Way, 2007; Vispoel, Hendrickson & Bleiler, 2000; Zheng, 2012). However, simulated databases have been used for most of the investigations. Empirical researches have mainly been conducted among university students; moreover, they were pilot studies with small samples.

The research presented here attempts to fill this gap. The main aim of the research is to examine the effectiveness of traditional linear testing in a comparison of linear testing modes among students in grades 1–8. The research mainly focuses on the technical characteristics of the adaptive testing process and the development, implementation and operation of computer adaptive testing (CAT) and considers its potential benefits among primary school age children.

THEORETICAL SOURCES FOR THE DISSERTATION

Adaptive tests operate according to a strict algorithm (Linacre, 2000; Magyar, 2013). At the beginning of the test, an initial item/module is selected from the pool. If preliminary information is available about the examinee, then it is used during the selection. If not, then one or more items are randomly selected from the item pool or its subset. The testing process usually starts with a medium-difficulty item. On an item-based test, a new item is selected after the first item is completed. If the answer is right, a more difficult item is presented; if not, an easier one follows. The program algorithm ensures that each subsequent item is matched to the individual's ability level. The items are delivered, scored and evaluated, and the program calculates if it is necessary to select a new item or if the test has ended. At the end of the testing, one can receive immediate feedback on the results achieved (Csapó, Molnár & R. Tóth, 2008; Eggen, 2004). CAT is thus dynamic, personalized and adapted to the level of the individual. The students do not begin with the same items, each student can receive different subsets of items, and they can receive different numbers of items from the item pool. The difficulty of the test is adjusted to the students' ability level as the test is being administered (Weiss, 2011).

A great variety of different forms of adaptive test models has been developed (van der Linden, 2008); they primarily differ with regard to the level at which the adaptivity occurs. Starting from item-based tests towards subtest-based multi-stage tests, all of them have the same basic structure in principle. Item-based adaptive tests have a number of advantages, but also disadvantages and limitations. With most item-based adaptive tests, there is no option for item review. Due to the random item administration, the previous item can provide information for the subsequent item, and the location of the items can also affect the answer; that is, the same item can be easier or more difficult, depending on its position in the test (Wainer & Kiely, 1987; Wainer, 2000; Linacre, 2000).

The problems that arise from the item-based CAT approach to testing may be overcome through the application of an alternative model, the multi-stage test (MST; Magyar, 2013), which combines characteristics of both the traditional linear and adaptive tests, because it adapts the items to the ability of the students like CAT and also provides an opportunity for the predetermination of the item order like PP tests (Armstrong, Jones, Koppel & Pashley, 2004; 2008; Molnár, 2013). In terms of their structure, they are halfway between the traditional linear and the item-based adaptive tests (Jodoin, Zenisky & Hambleton, 2006; Patsula, 1999; Zheng, 2012). During the test administration, short fixed subtests (modules) are provided in several stages, instead of items. A test consists of two modules at least. A stage comprises two or more modules, which differ in their difficulty levels. When the student finishes a module, his/her ability level is assessed; he/she is thus administered a more difficult or easier module in the next phase (Zenisky, Hambleton & Luecht, 2010).

The MST allows for a great variety in terms of the number of stages, modules and items on the test (Davis, 2005; Yan, von Davier & Lewis, 2014). The number of stages mainly influences the complexity of the test. The greater the number of stages, the greater the possibility of creating more possible routes and test versions. However, when the number of stages is increased, test administration becomes more complicated without the proportional increase of measurement precision (Armstrong et al., 2004; Hendrickson, 2007). Taking this into consideration, the most widespread forms are the 2-4-stage tests (Zenisky, Hambleton & Luecht, 2010).

The numbers of modules can have a variable number within a stage. Generally, a particular test starts with a single module, and thereafter modules at a particular stage number three to five (Patsula, 1999). However, to achieve the appropriate precision, three or four modules are sufficient at a certain stage (Armstrong et al, 2004).

During the testing process, students are administered easier or more difficult modules in the next phase on the basis of their performance. At the routing points, the algorithm used determines how the students are assigned to the various difficulty modules (Zenisky & Hambleton, 2004; Armstrong, 2002; Zenisky, Hambleton & Luecht, 2010).

The routing rules are closely related to the scoring of the modules and the complete test. As can be seen in the routing rules, when a student finishes a module, his/her ability level is estimated.

Often estimated ability levels are converted into number correct (NC) scores by the algorithm (Keng, 2008; Yan, von Davier & Lewis, 2014). Although NC scoring is sufficient for the scoring of the modules, it is not suitable for the whole test, as students are administered different test versions with various difficulties (Zenisky, Hambleton & Luecht, 2010). Therefore, the same methods are applied to score the whole test as are used for the item-based adaptive tests, using item response theory (Keng, 2008).

If we compare the multi-stage and item-based adaptive tests, the multi-stage tests have a number of advantages over the item-based tests (Magyar & Molnár, 2013). The modules can be designed and constructed in advance; greater control can therefore be ensured for test administration. Cross information among items can thus be eliminated (Hendrickson, 2007). Their use is particularly advantageous in the case of content restrictions (Hendrickson, 2007). Further important advantages include students enjoying opportunities for item review and correction (Zheng, 2012). As adaptivity is only achieved between modules, the test algorithm is not influenced and the students are motivated to score higher points (Vispoel, Hendrickson & Bleiler, 2000). Comparing the item-based tests, less administration and computer-based calculation are required (Hendrickson, 2007; Zheng, 2012).

However, Hendrickson (2007) emphasizes that multi-stage tests also have some shortcomings. More items are usually required to achieve the same precision as item-based tests. Test construction involves more work, as it calls for item effects to be examined in addition to the items themselves. With two-stage tests, the students' ability level may easily be estimated with greater error on the routing test. A further disadvantage is that the test can only terminate at the end of a certain module; the test is thus less flexible than item-based tests (Zheng, 2012). In spite of these disadvantages, the MST represents a balance with regard to accuracy, adaptivity, usability and control over the items (Zenisky, Hambleton & Luecht, 2010).

As long as there are several versions of a test, it is essential that the scores from the different versions be comparable to each other. This is particularly crucial in the case of longitudinal research or when the two testing modes are applied alternatively and in parallel (Way, Davis & Fitzpatrick, 2006; Paek, 2005). Professional testing standards (APA, 1986; AERA, APA & NCME, 1999; Wang, Wang, Jiao, Young, Brooks & Olson, 2008) also stress the importance of the comparability of scores achieved in different mediums.

The main focus of comparative studies is to compare the measurement precisions of the tests and to explore the impact of the transition to adaptive testing on the testing process (time and number of items) and on students' results.

Comparing adaptive and paper-based tests poses a great challenge (Wang and Kolen, 2001). As students are administered tailored tests, there may be differences in the content of the items, the situation, the difficulties of the items and their scoring. These factors can influence comparability significantly, and, as with the media effect, it is recommended that this be taken into account (Wang & Kolen, 2001; Kolen, 1999-2000). Wang and Kolen (2001) point out that in order for a CAT version to be comparable with its paper version there are strong limits on CAT as developers cannot take advantage of all the opportunities provided by computers during test development. However, today we are in a phase of transition from paper-based to computer-based testing, and there is a great need for comparability studies to ascertain trends; these researches are therefore especially justified (Wang and Kolen, 2001; Pásztor-Kovács, Magyar, Hülber, Pásztor & Tongori, 2013; Wan, Keng, McClarty & Davis, 2009).

Among the researches, many of them have investigated different constructions of MSTs. A variety of types have been applied depending on the content of the tests, the size of the item pool and other psychometric properties. In most cases, the testing process started with a medium-difficulty module and branched into 2–5 with the application of 2–6 stages. The increase of the numbers of modules regularly increased the measurement precision of the test. In most of the research, more than two stages were suggested because the incorrect classification of students can thus be minimized, as the differences can be eliminated at the latter stages. However, it is not recommended that too many stages be used, because the length of the test increases without the growth of measurement precision. Among the often used types are the 1-3 (Rotou et al, 2003), 1-2-3-4, 1-2-4 (Zheng, 2012), 1-3-3 (Keng, 2008), 1-2-2, 1-3-3, 1-2-3, 1-3-2 (Jodoin, Zenisky & Hambleton, 2006), 5-5-5-5-5-5 (Crotts et al, 2013) and 1-3-3-3-3 (Brossman & Guille, 2014) structure MSTs. In several studies, if the pool size was suitable, more equivalent test versions or modules were applied in order to enhance test security, which were administered in random order.

In the first comparability studies, classical test theories were applied (ANOVA analysis, comparing the means; Vispoel, Hendrickson & Bleiler, 2000; Olea et al., 2000); however, the use of item response theories later became common, in particular the comparison of item and test information. The main indicators of the measurement precision of a test are the reliability and the standard error (SE). In the case of simulation studies, an often used index is the correlation of true scores and estimated scores, the RMSE (Root Mean Square Error) and the AAD (Average Absolute Difference) indices, which show the differences between the true and estimated scores (Keng, 2008).

The studies confirm the assumption that more accurate ability estimates can be obtained with adaptive testing models compared to traditional linear test designs. The testing time is reduced, and fewer items are sufficient. Adaptive designs cover more information at every ability level, and the standard error is also significantly reduced. According to most studies, the measurement precision of the MST design is somewhat lower than that of the item-based or testlet-based adaptive design; however, the higher control during test development makes this test form beneficial. This is the main reason why this test construction is used most often during the transition period.

THE RATIONALE BEHIND THE EMPIRICAL STUDIES

AIMS

The main aim of the research is to examine the effectiveness of the adaptive testing method in comparison with linear test performance among students in grades 1–8 in large-scale measurements. Further aims are:

(1) To convert a previously paper-based test into an online form;
(2) To develop a linear and an adaptive test system that can be used in a classroom environment;
(3) To compare the measurement precision of adaptive and linear tests;
(4) To compare the ability levels within grade and at the individual level;
(5) To compare the rate of correct answers in the two kinds of test environment;
(6) To characterise the difficulty levels of the different items and modules administered with the adaptive method;
(7) To compare the amounts of information extracted from linear and adaptive tests as well as their standard errors.

RESEARCH QUESTIONS

Based on the outlined objectives, the following research questions were asked:

(1) Can paper-and-pencil-based tests be converted into an online test format?
(2) Is it possible to develop an adaptive testing system from the converted items that can be applied reliably in a classroom environment?
(3) Are adaptive tests more precise than linear tests?
(4) Are there any differences between the estimated scores on the adaptive and linear tests within grade or at the individual level?
(5) What is the proportion of correct answers achieved by the student in the two test environments?
(6) Which difficulty level items/subtests occur most often?
(7) What amount of measurement errors and information can be obtained from the two test environments?

Hypotheses

(1) The paper-based test can be converted into multi-stage adaptive test systems;

(2) Adaptive test systems can effectively and reliably be applied for the diagnostic measurement of 1st–8th graders;

(3) The adaptive system facilitates a more precise ability measurement;

(4) There are no significant differences between the calculated ability levels in the adaptive and linear testing methods;

(5) In the case of adaptive testing, the proportion of right answers is higher for the lower ability student in the lower ability section, and vice versa for the higher ability sections; namely, the proportion of right answers is lower for the high ability students on the adaptive test than on the linear version;

(6) The majority of students are at the average ability level; therefore, in the case of the adaptive testing method, the medium-level items/modules are administered the most frequently;

(7) In the case of adaptive testing, the calculated information is significantly higher at every ability level; in contrast, standard errors are significantly lower than with the linear testing method.

The process of the investigation

The investigation was conducted between 2012 and 2014 in several parts through large-scale tests and pilot measurements. Three pilot measurements and two large-scale measurements were conducted, for which the tests were assembled from different competence measurement tests. The first pilot measurement was carried out through an inductive reasoning test; in the second pilot measurement, problem-solving ability was investigated. The third measurement was conducted with a word reading skills measurement test system, which was tested first in a pilot measurement and then in a large-scale measurement.

During the two pilot measurements (inductive reasoning and problem-solving), the paper-based parameters were used for the composition of the adaptive test versions. In the case of word reading measurement, however, a separate large-scale measurement was conducted to calibrate the parameters of the items. The large sample measurement was carried out with these parameters. To ensure the comparison of the tests with each other, a linear test version was developed in each case. The adaptive test versions were multi-stage tests in different structures depending on the size and composition of the selected item pool.

Introducing the measurement instruments

The measurement tests that were used are suitable for measuring skills or skill areas; they play a major role in primary school students' skills and ability development. All three instruments were originally developed for paper-and-pencil-based testing, and they have been used for large-scale measurements a number of times. During these measurements, these tests have worked very reliably; the reliability index of each test was over 0.80 (Cronbach's alpha). The items were parametrized on the basis of these measurements. The difficulty indices of the test parameters covered the whole range of the ability scale, so these parameters were used during the small-scale measurements. Another reason for the choice of these tests was that they contained an appropriate number of items to develop the item bank for the adaptive tests. Another important aspect was that it was possible to convert the items into a computer format with only a slight modification, so the media effect did not significantly affect the validity of the tests.

During the conversion of the inductive reasoning and problem-solving ability tests, the originally paper-based parameters were used. In order to increase the measurement precision for the word reading test, online parametrization was carried out for all the items in the entire item pool, and this parametrized item pool was used for the development of the adaptive pilot and for the large-scale measurement. The 1-3-3-3 MST structure was used for the inductive reasoning test, and the 1-2-3 structure was applied for the problem-solving measurement test, as these structures were the most appropriate forms according to the literature. For the third study, in relation to the word reading skills measurement, a more complicated structure was preferred because of the large item pool size and the original test structure. Thus, in this case the 1-4-5-5 structure MST was chosen.

Criteria for assembling the samples

The research aims to investigate the possibilities for adaptive testing methods among primary school students. Therefore, given the options of the available tests, an attempt was made to involve the greatest range of the age groups studied. As was shown, the tests proved to be suitable within the broad age limits for ability measurements, so it was possible to extend the measurements for the upper and lower age range. In the measurements, a total of 8165 students took part among primary school grades 1–8. The inductive reasoning and problem-solving tests were used among 5th–8th graders. The word reading test was applied among the elementary school ages. A narrower age interval was chosen for the large-scale investigation: the 4th–5th graders were selected in order to eliminate the differences in age development.

Data collection

In all cases, the tests were prepared and administered with the eDia system. The students completed the tests in their own school through the schools' internet network. In each case, the measurement process took place for one lesson (45 minutes). The teachers received detailed measurement training with a detailed description of the measurement process. Each student had to complete two test versions. In the first phase, they were randomly administered a linear or adaptive test; in the second phase, reverse test administration occurred. Those who received the linear test in the first phase were provided an adaptive version, and vice versa. There was a minimum of two weeks' time and a maximum of four weeks left between the two phases. At the end of the measurement, the system provided feedback on the students' performance. In the case of the linear test, the students' results were indicated by the percentage they reached on the test; in the case of the adaptive tests, however, ability scores were computed by the program.

Data analysis

During the data analysis, the focus was on the examination of the technical operations of the tests, as the main aim of the research was to compare the items and their characteristics in the two types of test environments. The analyses were carried out based on classical test theory methods and with the application of item response theory. The classical test theory analyses were carried out with the SPSS program; the modern test theory analyses were performed with ConQuest. Item response theory made it possible to bring all of the items on one scale, called the ability scale, and this facilitated an independent analysis from the population on a probabilistic basis. The analyses were performed with the use of the partial credit model. The items were parametrized with the one-parameter Rasch model (Rasch, 1960). The estimated ability points were transformed to an average of 500 points and a standard deviation on a 100-point scale. Although both abilities and skills were estimated during the research, *ability scale* is the widespread term based on probability test theory, so the current term was not changed.

To determine the measurement accuracies of the tests, the reliability index, Cronbach's alpha, was used. Cronbach's alpha, however, can only be calculated when all the individuals participating in the testing process complete all the items, so there is no missing data. In the case of adaptive tests, however, the students complete only a subset of the item bank; hence Cronbach's alpha cannot be calculated. Therefore, the WLE person separation reliability was calculated to characterise the reliability of adaptive tests, which can be computed with item response theory and always provides a lower value than Cronbach's alpha (Linacre, 1997; Clauser & Linacre, 1999). Additional indicators of the measurement accuracy of the tests include the amounts of information and the standard error (Weiss, 2013), which were also calculated with the Rasch model. The test information curves characterise the test information by using the differences between the average-ability levels of the students and levels of difficulties of the test items that they complete. The amount of extracted information was considered maximum if the difficulty levels of the tests and the ability levels of the students were the same. The further away these values were from each other, the less information was obtained during the testing process.

Since each student completed both test versions, it was possible to make an individual-level comparison of the estimated ability levels and the percentage of correct answers. The relationships between the variables were examined with correlations. The significance levels of differences were determined with paired t-tests and analysis of variance (ANOVA). The effect size (Cohen's d; Cohen, 1988) was used to characterise the differences of the two tests.

Results and discussion

Based on the first pilot study, the reliabilities of the tests were suitable for measuring the inductive reasoning abilities of the young students. The person separation reliability of the adaptive version (0.85) was higher than the Cronbach's alpha indicator of the linear version (.83). The students' results on the adaptive and linear versions correlated highly (r=.82, p<.01); with regard to the grades, there were only significant differences between the two test forms among the results for the 8th graders. The lower-grade students' results did not differ significantly in the different test environments. Among the 17 test versions of the adaptive test, six of the routes occurred in the largest proportion; among those, the proportion of the average difficulty modules was the highest. Since the majority of students were of average ability, this corresponds to the expected rate. During the testing process, the students with the lower and higher abilities were clearly separated, and by the end of the test the students were divided into about the same proportion among the three ability zones.

The information extracted from the whole test was significantly higher, and a more accurate ability level measurement was achieved with the adaptive algorithm. Comparing the amount of the information extracted, the linear test provided an average 60 percent information rate; on the adaptive version, the average amount was 76 percent. Comparing the differences at individual levels, the difference was particularly significant in the low and high ability sections: nearly 34 percent in the lower section, and nearly 24 percent in the higher section. The standard error also decreased in the adaptive test algorithm.

Based on the second study, the reliabilities for both tests were acceptable; however, the adaptive version had a higher result (.83) than the linear version (.80), indicating a more accurate measurement precision of the adaptive system. The students' results in the two kinds of testing environment correlated highly (r=.71), and, according to the t-test, there were no significant differences between the results in the different testing modes (t=-.03, p=.98). Based on the grade-level analysis, there were also no significant differences between the results in the two systems. The study also compared the proportions of correct answers in the different testing modes. The comparison showed that, except for the eighth grade, each grade had a higher number of correct answers in the adaptive test environment than on the linear test. According to the ability levels, the number of correct answers increased with the rise of the skill level, but students below the average produced more correct answers on the adaptive test than on the linear test. For students of high ability, it showed the reverse; on the linear test, they produced a lower rate of correct answers. The six modules on the adaptive test provided an opportunity for a total of four different test versions. Slightly more than half of the students progressed on the route of the easy modules, and nearly half of the sample ended with the easy modules. Comparing the subtests at the individual level, the advantage of adaptivity was mainly manifested among the high ability students.

There was significantly more information available in the case of the adaptive version, and the measure of standard error was consistently higher for the linear test. The advantage of the adaptive test environment was significant in the higher ability range, as an average of 20–25 percent more information was obtained here than on the linear test.

Based on the word reading skills measurements, the system worked correctly during the pilot testing. In the case of the lower ability students, the typically easier clusters were administered during the testing process, while the more difficult ones were administered to the higher ability students; the amount of information obtained during the test thus gradually improved as each student received the appropriate difficulty level items in the latter phases. With regard to the last two modules, the module level did not change during the step from the third module to the fourth in the case of 31 students, which is only one-fifth of the students, so it was definitely appropriate to apply the fourth sections.

According to the results of the tests, a more accurate estimation was accomplished in the case of the adaptive system. The difficulty indices of the items covered the ability levels of the age group under investigation, so the test was able to estimate the levels of their ability correctly. Dimensions of the test closely correlated with each other; similarly, the two test results showed a strong correlation, meaning there were no significant differences between the results obtained by the students in the two kinds of testing environments. Comparing the rate of correct answers on the tests in the case of the lower ability students, there was a higher percentage of correct answers in the low ability zones; in the case of the higher levels, however, the number of correct answers was lower than it was for the linear test. This shows that the adaptive allocation provided more motivation for the weaker students and a challenge for students with high abilities.

As a significant number of students were of average ability, the medium difficulty section had the highest frequency during test administration. However, in the third stage and even in the fourth, the levels of difficulty changed for many students, thus justifying the need for five different difficulty-level modules.

During the measurement, the test information and the measured errors were compared, and in both cases both the test information and the measured error were favourable in the case of the adaptive test.

Confirmation of hypotheses

The aim of the thesis was to examine whether the transition from traditional linear testing to an adaptive method ensures higher measurement precision and to what extent can it be achieved. During the research the following hypotheses were raised and confirmed:

(1) The previously paper-based test can be converted into multi-stage adaptive test systems;

During the research, three, previously paper-based test systems were converted into online forms. In each case, the students' results in the two kinds of testing environment correlated highly, and according to the t-test there were no significant differences between the results in the different testing modes. The first hypothesis was confirmed, namely that the previously used paper-based systems can be converted into computerized adaptive versions.

(2) Adaptive test systems can be effectively and reliably applied to the diagnostic measurement of 1st–8th graders;

One indicator of the effectiveness and reliability of the test is the reliability index. Due to the fact that during an adaptive testing administration process, multiple versions were administered to the students, each student completed only a subset of the entire item pool. Therefore, Cronbach's α reliability index could not be used. Instead, in the case of an adaptive test, its extension, the person separation reliability index was used (WLE - Weighted Likelihood Estimate). The value of the person separation reliability index proved to be adequate for each of the adaptive tests, so the second hypothesis was confirmed: the developed adaptive tests are suitable for the diagnostic measurement of 1st–8th graders with respect to their reliability.

(3) The adaptive system makes a more precise ability measurement possible;

The measurement accuracy of the tests can be well characterised by the measurement error and the rate of the extracted information (Wang & Kolen, 2001; Wang, 2010; Molnár, 2013). The extent of the extracted information was characterised by the differences between the average skill level of the students and the level of item difficulties. The size of the extracted information was considered maximum if the difficulty level of the items and the students' ability level were the same. The bigger the difference was between them, the less was the rate of the information obtained during the testing process. In all three measurements, the rates of information from the adaptive test were significantly higher than on the linear test; therefore, a more accurate and precise measurement precision was achieved with the adaptive versions. Similarly, the computed errors were significantly lower in the case of adaptive tests, thus also proving that adaptive tests measure more accurately. As a result, the third hypothesis was also proved.

(4) There are no significant differences between the calculated ability levels in the adaptive and linear testing methods;

The differences between the estimated ability levels were examined in the case of the adaptive and the linear versions. In all the cases, the results showed high correlation coefficients, indicating that the two test versions classified the students similarly. According to the results of the t-tests, there was only one grade where the differences between the results of the tests were significant, suggesting that overall there was no significant difference between the two ability estimates. The fourth hypothesis was partly confirmed.

(5) In the case of adaptive testing in the lower ability section, the proportion of right answers is higher for the lower ability student, and vice versa for the higher ability sections; namely, the proportion of right answers is lower for the high ability students on the adaptive test than on the linear version;

The number of correct responses increased with the increase of ability levels on both test versions. However, in the case of the adaptive version, the rate of growth was different than those found during linear testing. During the adaptive testing, the proportion of correct answers at the low range ability level significantly increased, a result which can be explained by the fact that the lower ability students were administered easier items that could be completed more easily, so they managed a greater proportion of correct answers and felt a greater sense of achievement. In the high ability zone, it happened the other way around; the higher ability students were exposed to more demanding items. Thus, fewer correct answers were achieved, so the test was more challenging for them. In the case of inductive reasoning measurement, the proportion of correct answers was higher at every ability level. The fifth hypothesis was partly confirmed.

(6) The majority of students are at the average ability level; therefore, in the case of the adaptive testing method, the medium-level items/modules are administered the most frequently;

The distributed routes were examined during the adaptive testing, and in all cases the medium modules occurred the most frequently. Since the majority of students were at the average ability level, most students passed along the medium difficulty routes. To increase the security of the test, it is advisable to increase the number of medium difficulty modules. The sixth hypothesis was proved.

(7) In the case of adaptive testing, the calculated information is significantly higher at every ability level; in contrast, the standard errors are significantly lower than in the linear testing method.

The amount of information obtained from tests was analysed for each study. More information was acquired from the adaptive test data than from that of the linear tests in all of the cases regardless of data measurement. However, the extent of these differences varied at different ability levels. In the measurement of inductive reasoning, the extent of the information extracted in the high and low ability ranges was significantly greater than for the linear test. In the problem-solving ability measurement, particularly in the case of high ability ranges, the amount of information extracted was significantly higher. In the measurement of word reading ability skills, the differences between the extent of the information from the adaptive and the linear testing were more consistent; about the same amount of information was thus extracted across the full ability scale at all skill levels. The degree of measurement error was the same in all of the cases. To sum up, the seventh hypothesis was partially confirmed: the size of the measurement error and the amount of extractable information may be different at the different ability levels depending on the sample size. According to the results of the study, in the case of small samples, particularly in the low and high ability ranges, the amount of available information is significantly higher from the adaptive tests; in a larger sample, it becomes more balanced, and it is possible to consistently obtain more information from the adaptive test than from the linear version at all levels of the ability range. The degree of measurement errors thus indicated a similar degree.

Conclusions

The emerging needs of assessment and evaluation in the 21st century clearly indicate the way towards the development of computerized testing. Computerized tests offer a number of new opportunities for the measurement of abilities; with their help immediate evaluation has become possible, new and innovative item types can be worked out, and new ability areas can be measured accurately and efficiently. The conversion of paper-based tests into computerized versions can be achieved at a number of levels. Currently, the most innovative form is computerized adaptive testing. In the case of adaptive testing technology, the items or subtests are administered from a pre-measured and parametrized item pool in such a way that each student is administered the item or subtest which best matches his/her ability level. This testing mode facilitates a more accurate and efficient measurement compared to the traditional, linear testing method. Since the students are administered items tailored to their own ability level, the test is equally challenging for them from the beginning to the end, so each item on the test equally contributes to estimating the individual's ability level; a more accurate ability level measurement thus becomes possible.

There are several types of adaptive tests; one of the most preferred types is the multi-stage adaptive test structure, in which modules are administered instead of items, modules actually being fixed short tests with different levels of difficulty. The test type combines the properties of fully adaptive tests and traditional linear tests, as the difficulties of the items are adjusted to the students' ability level and the option of determining item order beforehand is also provided. The modules can be planned and developed in advance, thus allowing greater control over the administration of the test, so one can avoid subsequent items providing information on each other. Another important advantage is that the students have the opportunity for item review and correction. Since adaptivity only takes place between the modules, it does not jeopardize the test algorithm and helps students to achieve the highest possible score. Compared to the item-based adaptive tests, they demand much less administration and computer calculation.

In the transition from paper-based tests to adaptive tests, it is very important to examine whether the transition ensures the expected level of improvement in measurement precision and more effective ability measurement. According to the relevant international studies, adaptive tests have higher reliability than linear tests and the amount of extractable information is also higher. Nevertheless, the measurement error is lower than with the linear tests; therefore, greater measurement accuracy is possible. The measurement precision of multi-stage tests is slightly lower than that of item- or testlet-based tests; however, the greater administrative control makes this test type advantageous.

During the transition from traditional testing to adaptive testing, comparative analyses of the two forms of tests have been conducted in many cases. However, the results of international studies are primarily based on simulated databases; only a few of them have been conducted with real samples, and these involved college or university students.

The purpose of the research was to examine the possibilities for the conversion of paper-based testing to the adaptive testing method through empirical studies among primary school students. During the research, several ability areas and different versions of adaptive structures were investigated through a comparison of their effectiveness and measurement precisions against the traditional linear testing environment.

The study was conducted in several phases. Three pilot measurements and two large-scale measurements were carried out involving three different ability areas. The samples for the study were 1st–8th graders. During the pilot studies, the students' inductive thinking, problem-solving ability and word reading ability were measured, followed by two large-scale data collection efforts, in which word reading abilities were also investigated. The tests under investigation were paper-based tests converted into online form in the eDia system. The data collection was done in the students' own classroom during one lesson time, which is 45 minutes. To ensure the individual-level comparison, each student completed both test versions, conventional linear and adaptive. As adaptive tests are based on item response theory, it was this theory, i.e. the one-parameter Rasch model, that was used during the data analysis.

In line with the relevant international research, the analysis compared test indicators and estimated ability levels in adaptive and linear test environments. The study also investigated the characterisation of the difficulty levels of the adaptive versions and the proportions of correct responses. Further, it compared the amounts of information extracted and measurement errors in both testing methods.

The results confirmed the hypotheses, and, in accordance with the literature, the research shows that paper-based tests can be converted into adaptive forms and can be used effectively among primary school students to estimate their level of ability. In the case of adaptive testing technology, lower ability students completed more correct answers, so the test was more motivating for them; for students with higher level skills, it was more challenging for them to complete the tests. The adaptive tests make a more accurate ability measurement possible, have a higher reliability, and with regard to the entire sample, significantly more information can be extracted from them than from the conventional linear tests. However, depending on the sample size, the degree of information obtained may be different on different ability levels. For smaller samples, mainly in the low and high ability ranges, the information extracted is significantly higher in the case of the adaptive test; in larger samples, it becomes balanced and nearly the same proportion of information can be obtained from the test in all the ability ranges. The size of the measurement error also occurs similarly, and the size of the estimated error changes again with the size of the sample.

The uniqueness of the research is that unlike most studies that use a simulated database, here empirical data were used during the comparison; moreover, the use of the same sample also made student-level comparisons possible. The studies were conducted among 6–14-year-old students, a feature which is also unique in the research on adaptive testing. The results confirmed those of international simulation experiments, namely, that considerable measurement precision can be achieved using an adaptive test algorithm compared with conventional linear tests.

The limitations of the findings are that the possibility of switching from the linear to the adaptive testing method was only investigated in three ability areas during the research. The item pools that were used differed with regard to item size and type. These characteristics may influence the amount of available information; additional research is therefore needed using a variety of item pools.

The practical significance of the research is that it has developed tests that can be used in a classroom setting and provide immediate feedback both to students and teachers. The students' ability estimate became more accurate, a development which can significantly influence students' achievement on criterion-referenced tests in particular. During the testing process, different items are administered to students; the security of the test can thus be increased. As the items are administered from a parametrized item bank, the students' results can be characterised on the same scale, the ability scale. It is a likely indicator of how the students would perform on other items even if they did not complete all of them.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessment (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: Author.

Armstrong, R. D. (2002). *Routing rules for Multiple-Form Structures.* (Computerized Testing Report 02-08). Newtown: Law School Admission Council.

Armstrong, R. D., Jones, D. H., Koppel, N. B., & Pashley, P. J. (2004). Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement*, *28,* 147–164.

Al-A'ali, M. (2007). Implementation of an improved adaptive testing theory. *Educational Technology & Society*, *10*(4), 80–94.

Beller, M. (2013). Technologies in large-scale assessments: New directions, challenges, and opportunities. In: *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research.* Netherlands: Springer, 25–45.

Brossman, B. G., & Guille, R. A. (2014). A Comparison of multi-stage and linear test designs for medium-size licensure and certification examinations. *Journal of Computerized Adaptive Testing*, *2*(2), 18–36.

Clauser B. & Linacre J.M. (1999). Relating Cronbach and Rasch Reliabilities. *Rasch Measurement Transactions. 13*(2), 696–697.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Crotts, K. M., Zenisky, A. L., Sireci, S. G. & Li, X. (2013). Estimating measurement precision in reduced-length multi-stage adaptive testing. *Journal of Computerized Adaptive Testing. 1*(4), 67–87.

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

Csapó Benő, Molnár Gyöngyvér & R. Tóth Krisztina (2008). A papír alapú tesztektől a számítógépes adaptív tesztelésig: a pedagógiai mérés-értékelés technikájának fejlődési tendenciái. *Iskolakultúra, 18*(3–4), 3–16.

Davis, S. L. (2005). *Exploring a new methodology for setting performance level standards with computerized adaptive tests*. 35th Annual National Conference on Large-Scale Assessment. Texas, TX: San Antonio.

Eggen, T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing*. Netherlands: Citogroep Arnhem.

Eggen, T. J. H. M. (2007). Choices in CAT models in the context of educational testing. In: Weiss, D. J. (ed.): *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* http://publicdocs.iacat.org/cat2010/cat07eggen.pdf. Retrieved: 12.12.2013.

Eggen, T. J. H. M. & Straemans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories*. Educational and Psychological Measurement, 60*(5), 713–734.

Frey, A. & Seitz, N. N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, *35***(**2–3), 89–94.

Frey, A., Seitz, N. N. & Kröhne, U. (2011). Reporting differentiated literacy results in PISA by using multidimensional adaptive testing. In: Prenzel, M., Kobarg, M., Schöps, K. & Rönnebeck, S. (ed.): *Research in the context of the Programme for International Student Assessment*. Berlin: Springer. 103–133.

Greiff, S., Wüstenberg, S. & Funke, J. (2012). Complex Problem Solving. More than reasoning? *Intelligence*, *40*(1–14).

Guille, R. A., Becker, K. A., Zhu, R. X., Zhang, Y., Song, H., & Sun, L. (2011). *Comparison of asymmetric early termination MST with linear testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Hambleton, R. K. & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass–fail decisions. *Applied Measurement in Education*, *19*(3), 221–239.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*(2), 44–52.

Jodoin, M., Zenisky A. & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education. 19*(3), 203–220.

Keng, L. (2008). *A Comparison of the performance of testlet-based computer adaptive tests and multistage tests.* Austin: The University of Texas.

Kingsbury, G. G. & Hauser, C. (2004). *Computerized adaptive testing and the No Child Left Behind.* Presented at the Annual Meeting of the American Educational Research Association. San Diego, CA.

Kolen, M. J. (1999-2000). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment*, *6*(2), 73–96.

Linacre, J. M. (1997): Kr-20/Cronbach alpha or Rasch reliability: Which tells the truth? *Rasch Measurement Transactions*, *11*(3), nos. 580–581.

Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come.* MESA Psychometric Laboratory, University of Chicago.

Molnár Gyöngyvér (2010). Technológia-alapú mérés-értékelés hazai és nemzetközi implementációi. *Iskolakultúra*, *20*(7–8), 22–34.

Molnár Gyöngyvér (2013). *A Rasch modell alkalmazási lehetőségei az empirikus kutatások gyakorlatában*. Budapest: Gondolat Kiadó.

Olea, J., Revuelta, J., Ximénez, M.C. & Abad, F. J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. *Psicológica*, *21*(1), 157–173.

Paek, P. (2005): *Recent trends in comparability studies*. PEM Research Report 05–05.

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing*. Electronic Doctoral Dissertations for UMass Amherst. Paper AAI9950199.

Pyper, A. & Lilley, M. (2010). *A comparison between the flexilevel and conventional approaches to objective testing.* CAA Conference. University of Hertfordshire.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen.

Rotou, O., Patsula, L., Manfred, S. & Rizavi, S. (2003). *Comparison of multi-stage tests with computerized adaptive and paper and pencil tests.* Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME). Chicago, IL.

Scheuermann, F. & Björnsson, J. (2009, eds.). *The Transition to Computer-Based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing.* Luxemburg: Office for Official Publications of the European Communities.

Scheuermann, F. & Pereira, G. A. (2008, eds.). *Towards a research agenda on computer-based assessment.* Office for Official Publications of the European Communities, Luxembourg.

Thompson, N. A. & Prometric, T. (2007). A practitioner’s guide for variable-length computerized classification testing. *Practical Assessment Research and Evaluation*, *12*(1), 1–13.

Thompson, T. & Way, D. (2007). *Investigating CAT designs to achieve comparability with a paper test.* Presented at the Applications and Issues Paper Session. Minneapolis, MN.

van der Linden, W. J. (2008). Some new developments in adaptive testing technology. *Zeitschrift für Psychologie*, *216*(1), 3–11.

Vispoel, W. P., Hendrickson A. B. & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests. Psychometric and attitudinal results. *Journal of Educational Measurement*, *37*(1), 21–38.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd Edition). Erlbaum, Hillsdale, NJ.

Wainer, H. & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185–201.

Wan, L. Keng, L., McClarty, K. & Davis, L. (2009). Methods of comparability studies for computerized and paper-based tests. *Test, measurements and research services bulletin, 12*(10), 1–4.

Wang, H. (2010). Comparability of computerized adaptive and paper-pencil tests. *Test, measurements and research services bulletin*, *13*(1), 1–7.

Wang, T. & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38*(1), 19-49.

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

Way, W. D., Davis, L. L. & Fitzpatrick, S. (2006). *Practical questions in introducing computerized adaptive testing for K–12 assessments*. San Antonio: Pearson.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1–27.

Weiss, D. J. (2013). Item banking, test development, and test delivery. In Geisinger, K. F. (eds.): *The APA handbook on testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology.* Washington DC.: American Psychological Association, 185–200.

Yan, D., von Davier, A. A. & Lewis, C. (2014). *Computerized multistage testing: Theory and applications.* New York: CRC Press,.

Zenisky, A., Hambleton, R. K. & Luecht, R. M. (2010). Multistage testing: Issues, designs and research. In: der Linden, W. J. & Glas, C. A. W. (eds.): *Elements of adaptive testing.* New York: Springer, 355–372.

Zheng, Y. (2012). Multistage Adaptive Testing for a Large-scale Classification Test: Design, Heuristic Assembly, and Comparison with Other Testing Modes, ACT research report series.

Related Publications

Gyöngyvér Molnár & Andrea Magyar (2015). *Comparing the measurement effectiveness of computerised adaptive testing and fixed item testing*. 16. EARLI Conference. Limassol, Cyprus, August 25-29. Accepted abstract.

Molnár Gyöngyvér & Magyar Andrea (2015). A számítógép alapú tesztelés elfogadottsága pedagógusok és diákok körében. *Magyar Pedagógia*. *115*(1), 49–66.

Magyar Andrea & Molnár Gyöngyvér (2014). A szóolvasási készség adaptív mérését lehetővé tevő online tesztrendszer kidolgozása. *Magyar Pedagógia*, *114*(4), 259–279.

Magyar Andrea, Molnár Gyöngyvér, Pásztor Attila, Pásztor-Kovács Anita & Pluhár Zsuzsa (2015). *21. században elvárt képességek számítógép alapú mérése.* XIV. Országos Neveléstudományi Konferencia, Debrecen, November 6–8.

Magyar Andrea & Molnár Gyöngyvér (2014). *A szóolvasási készség személyre szabott diagnosztikus mérését megvalósító online tesztrendszer kidolgozása*. XIV. Országos Neveléstudományi Konferencia, Debrecen, November 6–8.

Magyar Andrea & Szili Katalin (2014). *Application of Computer-Based Test to the Assessment of Reading Skills among Young Children.* VII. EARLI SIG 1: Assessment and Evaluation Conference. Madrid, August 27–29.

Magyar Andrea (2014). Adaptív tesztek készítésének folyamata. *Iskolakultúra*. *24*(4), 26–33.

Andrea Magyar (2014). *Problem Solving Competence Assessment with Computerized Adaptive Testing and Fixed Item Testing among Young Children*. XVIII. JURE Konferencia. Nicosia, Cyprus, June 30–July 4.

Magyar Andrea & Szili Katalin (2014). *Computer-based assessment of word reading skills*. XII. Pedagógiai Értékelési Konferencia, Szeged, May 1–3.

Magyar Andrea (2014). *Szóolvasási készséget mérő adaptív tesztelésre alkalmas feladatbank fejlesztése.* VI. Oktatás-Informatika Konferencia. Budapest, February 7–8.

Magyar Andrea (2014). *Szóolvasási készséget mérő adaptív tesztelésre alkalmas feladatbank fejlesztése.* VI. Oktatás-informatika Konferencia tanulmánykötete. 404–412. http://www.eltereader.hu/media/2014/03/VI_OKTINF_Tanulmanykotet_READER.pdf

Pásztor-Kovács Anita, Magyar Andrea, Hülber László, Pásztor Attila & Tongori Ágota (2013). Áttérés online tesztelésre – a mérés-értékelés új dimenziói. *Iskolakultúra*. *23*(11), 86–100.

Magyar Andrea (2014). *A problémamegoldó gondolkodás vizsgálata adaptív tesztek alkalmazásával.* A VIII. Kiss Árpád Emlékkonferencia előadásainak szerkesztett változata Tartalmi összefoglalók. 211–221.

Magyar Andrea & Molnár Gyöngyvér (2013). Adaptív és rögzített formátumú tesztek alkalmazásának összehasonlító hatékonyságvizsgálata. *Magyar Pedagógia*, *113*(3), 181–193.

Szili Katalin & Magyar Andrea (2013). *A szóolvasó készség számítógép alapú mérése*. Beszédkutatás Konferencia. Budapest, November 14–15.

Magyar Andrea (2013). *Különböző típusú adaptív tesztek hatékonyságának összehasonlítása*. XIII. Országos Neveléstudományi Konferencia. Eger, November 7–9.

Magyar Andrea (2013). *Problémamegoldó gondolkodás vizsgálata adaptív tesztekkel*. XIII. Országos Neveléstudományi Konferencia. Eger, November 7–9.

Magyar Andrea (2013). *Problémamegoldó gondolkodás vizsgálata adaptív és lineáris tesztekkel.* VIII. Kiss Árpád Emlékkonferencia. Debrecen, September 6–7.

Andrea Magyar (2013). *Comparative Study on Computerized Adaptive Testing and Fixed Item Testing.* 5th Szeged Workshop on Educational Evaluation, Szeged, April 15–16.

Magyar Andrea (2013). Többszakaszos adaptív tesztek felépítése, működése. *Oktatás-Informatika*, 1-2. http://www.oktatas-informatika.hu/2013/11/magyar-andrea-tobbszakaszos-adaptiv-tesztek-felepitese-mukodese. Letöltés ideje: 2014.06.20.

Magyar Andrea (2013). *Többszakaszos adaptív tesztek gyakorlati alkalmazása. XI. Pedagógiai Értékelési Konferencia*, Szeged, April 11–13.

Magyar Andrea (2012). Számítógépes adaptív tesztelés. *Iskolakultúra*, 22(6), 52–60.

Magyar Andrea (2012). *Számítógép alapú adaptív tesztek és fix tesztek összehasonlító vizsgálata.* XII. Országos Neveléstudományi Konferencia. Budapest, November 8–10.

Andrea Magyar (2012). *Comparative Studies on Computerized Adaptive Testing*. X. Pedagógiai Értékelési Konferencia, Szeged, April 26–28.