EÖTVÖS LORÁND UNIVERSITY
FACULTY OF EDUCATION AND PSYCHOLOGY

Theses of the Doctoral Dissertation

Judit Takácsné Kárász

Development of adaptive testing algorithms using National Assessment of Basic Competencies data

DOI: 10.15476/ELTE.2024.139

Doctoral School of Education
Head of the Doctoral School: Dr. habil. Anikó Zsolnai, Professor

Instruction-Learning-Inequality Module
Module Manager: Dr. Habil. Sándor Lénárd, Associate Professor

Supervisors:
Dr. Habil. István Nahalka CSc, ret. university Associate Professor
Dr. Habil. László Széll Krisztián, Associate Professor

Budapest, 2024

Contents

1. Introduction 2
2. Measurement Theory Background 4
3. Aims and questions of the research 6
4. Research methodology 7
5. Results 8
5.1. From paper-and-pencil to computer-based testing – mode effect study 8
5.2. From linear to adaptive testing - the role of open-ended items 10
5.3. Adaptive testing design – theoretical optimum 11
5.4. Comparison of possible adaptive strategies based on accuracy and reliability 12
6. Summary, limitations and directions for further research 13
References 15
Related publications 19
Non-related publications 20
Conference presentations on the topic 21
Off-topic conference presentations 23
Appendices 26

1. Introduction

In the 1980s, with the introduction of the *New Public Management* model, the delivery of public services, including education, became more decentralised in many countries (Hood, 1991). At the same time, tools for measuring the effectiveness and efficiency of the business world and accountability systems were also developed, so that after the turn of the millennium, a series of international student assessments began to compare the effectiveness of countries or education systems across different goals, domains and age groups.

The OECD PISA (OECD, 2023) measures the reading, mathematical and science literacy of 15-year-old students every three years since 2000. The purpose of the assessment is to examine whether students have the skills needed to learn independently in the world of work. Since 2001, the IEA PIRLS (Mullis & Martin, 2019) has tested the reading proficiency of 4th graders every five years at the end of reading instruction, while

the IEA TIMSS (Mullis et al., 2021) assess mathematics and science in 4th and 8th grade every four years since 1995, according to the curriculum.

Since 2010, these international student assessments have gradually moved towards computer-based and then adaptive testing. PISA in 2015 (OECD, 2017) and TIMSS in 2019 (Mullis & Martin, 2017) were mainly administered on computer, but the paper-and-pencil tests with bridge items only could also be choose to calculate trends. PISA 2018 measured the then highlighted domain of reading with a multistage adaptive test (OECD, 2019). PIRLS 2021 simultaneously introduced the computer-based assessment developed for TIMSS and the group adaptive assessment design, which handles different objectives and levels of difficulty (digitalPIRLS, ePIRLS, PIRLS Literacy). This design was also used by TIMSS 2023 (Mullis et al., 2021).

The National Assessment of Basic Competencies (NABC) in Hungary, which was developed on the basis of the PISA measurement methodology and has been administered since 2001, is organised by the Educational Authority. It measures students' reading and mathematics performance every year, initially in the 6th and 10th grades and later supplemented with the 8th grade (Balázsi et al., 2014). The aim of the assessment is to provide educational institutions, administrators and educational policy makers with objective performance indicators and to spread the international measurement culture (Csíkos & Vidákovich, 2012). At the beginning of the doctoral research, NABC was still a paper-and-pencil test, but in 2022 it was organised as a computer-based assessment (Balázsi et al., 2021). By 2024, science and digital culture were added to the domains, and the assessment interface merged the previously separate language assessments. Measured grades were also extended to 4–11 (Oktatási Hivatal, 2022). This shift to computer based testing can pave the way for further development before moving to an adaptive testing method.

In adaptive measurement (Magyar, 2012), the tasks completed by the student are scored during the test and, based on the estimated ability score using item response theory models (IRT), the next test section that best matches the difficulty level is assigned. In the case of multistage adaptive testing (MST), item groups are administered; in the case of computerised adaptive testing (CAT), the test is administered after each item. As a result, individual test paths are created and the corresponding items result in a more accurate assessment of ability and/or a shorter test. The development of such a test is costly in all respects, but the first stage of development can be well prepared with simulation tests (Thompson & Weiss, 2011).

The motivation and objective of my research is to prepare the ground for computerised adaptive testing for NABC in the domain of mathematics, through such preliminary studies, using the large amount of empirical data accumulated during the long history of the assessment. In the following, I will briefly summarise the measurement theory background of the dissertation. After the research questions, I will present my findings on the transitions from paper-and-pencil to computer-based to adaptive testing. The individual results also form a bridge between the data from the paper-and-pencil tests and the findings regarding the hypothetical adaptive testing. My study is a novelty in the Hungarian literature, a research with similar aim has been conducted previously in the case of multistage adaptive testing (Magyar & Molnár, 2015).

2. Measurement Theory Background

One of the indicators of test goodness is how accurately the measurement estimates the phenomenon (reliability), within which we can examine how well the individual items work together, or how successful the combination of items is in discriminating between test takers (internal consistency) (Nagybányai-Nagy, 2006). There are several measures of internal consistency, the three best known are the KR-20 (Kuder & Richardson, 1937) and Cronbach's alpha (Cronbach, 1947, 1951) based on classical test theory, and the person separation reliability (Wright & Masters, 1982) applicable to the Rasch model in modern test theory. Therefore, when simulation results of different tests are compared, or when the aim is to develop a test with adequate accuracy, one of these measures is usually used. Although Cronbach's alpha is the most commonly used, it cannot be used in adaptive testing because the relationship between certain items cannot be calculated due to the different response paths.

CAT (Weiss & Kingsbury, 1984) is a combination of adaptive testing, IRT methods and interactive computer-based survey administration. After an initial item, the computer estimates the respondent's ability based on all previous responses, and then selects the next item that best matches the estimated value. This cycle of testing continues until a termination criterion (stopping rule) is met. The CAT thus has six structural elements:

1) *IRT model*, according to which the characteristics of the items and the ability of the respondent are linked by an equation describing the probability of correct answer. On the basis of the responses, the parameters of the items and the respondent's ability can

be estimated. In the one-parameter Rasch model (Rasch, 1960), the items are distinguished from each other by their difficulty. In the two-parameter model the items also differ in their discriminatory power (slope), and the three-parameter model takes into account the fact that respondents with low ability development may tend to answer randomly to more difficult tasks (DuToit, 2003). The NABC uses the three-parameter IRT model.

2) *Item bank*. In the case of mastery tests (pass or fail), the difficulty of the items should primarily cover the area around the cut-off point; in the case of performance tests, they should cover the whole range of ability scale, and it is advisable if their discrimination (i.e. slope) is high. The size of the item bank depends on the size of the expected sample and can range from a few tens to a few hundred items (Magyar, 2014; Weiss & Kingsbury, 1984). In the case of larger or more frequently used assessments, some of the items will become known to many respondents, which may jeopardise the security of the measure.

3) *Entry level*. The starting point (initial ability estimation) may be the same for the entire sample population, but if some prior information is available, it can also be personalised.

4) *Item selection procedure*. The most common methods are maximum information, selection according to the difficulty closest to the estimated ability point, and the Bayesian approach, which usually produce very similar results. Studies on item selection procedures usually compare different methods according to the number of items needed to complete the test and/or the accuracy of the ability estimation, usually using simulation methods (Ito & Segall, 2013).

5) *Ability estimation*. Based on the current response pattern, the ability and possibly the confidence interval of the ability are estimated. Maximum likelihood and Bayesian estimation methods, or a combination of these, are used for estimation.

6) *Stopping rule*. According to the objectives of the test, one of several termination criteria may be a condition, for example a certain number of items has been answered; the standard error of the ability estimation falls below a certain level (an equally accurate ability estimate for all respondents); based on the confidence interval of the ability estimation, the test taker can be classified into a certain performance level (classification); the maximum time allowed for answering the questions has expired.

CAT has several advantages over linear tests. Tests are expected to be significantly shorter, and ability estimation is more accurate at the edges of the scale

(Weiss, 2011). A question with a difficulty level that matches the ability level can increase the intrinsic motivation to complete the test, although this is more likely to be experienced at the lower end of the ability scale if the test taker is informed about how the CAT works (Wise, 2014). The disadvantage of CAT may be that there is no opportunity to go back and improve on previous items, which may increase stress, but in their meta-analysis Akhtar et al. (2023) found no clear result for either increased motivation or increased stress.

Computerised measurement is required to carry out CAT on a large number of people, which means that the difference between the paper-and-pencil test and its computerised version and so the possibility of mode effect are raised (Buerger et al., 2019). The difference may be reflected in the difference between the constructs measured, in the systematic deviation of the performance scores, or in the characteristics of the text responses. A further difference may be caused by the different scoring of omitted and not reached items.

Because of the immediate scoring, open-ended items that require independent text creation and qualified coders cannot be used in the CAT, or at least they do not participate in ability estimation and item selection during the test. If these items measure different aspects of the phenomenon than the closed, automatically scored items, this also leads to a difference between linear and adaptive tests.

3. Research aims and questions

In my research I am looking for answers to the following main and sub questions.

1) Can the data from the OKM paper-and-pencil assessments be used in a relevant way for the design of the computerised adaptive assessment?

a. What is the effect of the computer-based testing environment on the test result? Should mode effect be expected, and if so, how should it be handled? (Fishbein et al., 2018)
b. Even if the open-ended items are omitted, does the assessment framework of the NABC remain the same? If only closed items were used, what would be the difference in students' ability estimation?

2) With the same measurement accuracy as the original assessment, what is the shortest test length achievable with adaptive testing? (Weiss, 2011)

3) Based on the data from the paper-and-pencil tests of the NABC, which adaptive testing elements are likely to make the measurement goal (in the domain of mathematics) more successful? (Thompson & Weiss, 2011)

a. With the same number of items as in the paper-and-pencil test, does the standard error of the student's ability estimation decrease in the case of the adaptive testing?
b. Do the simulations imitating the computerised adaptive test on the basis of NABC data confirm that it is possible to determine students' ability scores/performance in a significantly shorter time (with fewer items) with an accuracy equal to the accuracy of the paper-and-pencil test?
c. Which stopping rules correspond to which measurement objectives in relation to adaptive NABC?
d. Is there a hierarchy of stopping rules, i.e. are there strong criteria whose fulfilment requires the fulfilment of weaker conditions?
e. Does the student's performance change after the first 5–10–15–20 questions? With a test length of 5–10 questions, what performance estimates can be expected?

4. Research methodology

My research is a preliminary study to inform the next stage of developing the National Assessment of Basic Competencies, an existing assessment system, and accordingly applied research, in terms of its methodology it is primarily quantitative. I conducted the mode effect study (1a.) with a systematic literature review, which is a qualitative methodological study. The omission of open-ended items (1b.) is a quantitative, empirical study in which I used IRT ability estimation, correlation and cross tabulation analyses. The theoretical investigation of computer adaptive testing technology (2) is a qualitative mathematical derivation in relation to the standard error of estimating students' ability. The comparison of different item selection and ability estimation procedures are quantitative and empirical studies (3a.-e.), which I investigated with hybrid and Monte Carlo simulations.

Questions 1a.-b. have been examined in the fields of mathematics, reading and science. The similar test structure gives greater validity to the results in the field of mathematics according to the principle of triangulation. The examination of question 2 is independent of the domain, questions 3a.-e. are focused specifically on the mathematics questions.

In the case of adaptive testing, simulation methods are techniques based on high computer capacity and formal mathematical equations of IRT models. The point is that the computer calculates the probability of a correct answer based on the applied IRT model, using the student's theoretical ability and the well-known item parameters, and compares it with a random number between 0 and 1 to simulate the correct or incorrect answer. The ability score is estimated based on the simulated responses, and the item selection method selects the next item according to this estimate. The advantage of simulation over real data collection is that it allows for a larger number of sample and the comparison of many different conditions (e.g., Şahin & Weiss, 2015). Simulations can be grouped according to the extent to which they use real data. *Monte Carlo simulations* (Kehl, 2012) use data that are entirely generated by a random number generator. In *hybrid simulations*, some of the responses are real and some are generated based on the IRT model. We can speak of a *post-hoc simulation* when the real answer is available for each item and each respondent (Sari, 2020), as they were completed as a linear test.

The simulation of the CAT was carried out using the open source program package catR (Magis et al., 2017b), which operates in the R (R Core Team, 2016) environment and is free to use and programmable, i.e. adaptable to research.

5. Results

5.1. From paper-and-pencil to computer-based testing – mode effect study

In the case of NABC, there is no publication available on the study of mode effect, so I investigated this question by a systematic literature review (Rother, 2007) on digitalisation of the PISA, PIRLS and TIMSS assessments, including Hungarian and international scientific publications and assessment documents related to mode effect. I followed the PRISMA guidelines (Page et al., 2021) recommended for systematic reviews and meta-analyses, i.e. the aim of the database search is to identify and synthesise as many and as relevant sources as possible, based on well-defined inclusion and exclusion criteria. I conducted the search on 2 December 2021.

I conducted the domestic search using the MATARKA, MTMT and Arcanum Digital Library databases and the Educational Authority website, and the international search using the EBSCO, ERIC, JSTOR, ProQuest, Science Direct and Web of Science

databases, as well as the OECD and IEA websites. All peer-reviewed empirical quantitative articles, book chapters or studies published after 2010, in English or Hungarian, which used the original assessment data or data from own study closely related to it, and which specifically aimed to investigate the mode effect, were accepted. The domestic search found 2 relevant sources out of 375. Both are a summary of an assessment cycle for which I found 4 other similar documents. The documents give brief information about the introduction of computer-based assessment and mention the possible deviation of performance resulting from it. The international search yielded 1,262 items. 20% of the screened texts and the assess of the full texts was also done by a secondary coder. Based on Cohen's kappa, there was at least significant agreement between the coders. In the end, 8 items met the criteria (Appendix 1), to which 24 measurement documents were added. Among the measurement documents, those from PISA 2015 and TIMSS 2019 are relevant, PISA 2018 does not provide any additional information on mode effect, and in the case of PIRLS, there was no computer-based assessment or information on its preparation in the period considered.

The investigation of mode effect in PISA was conducted during the field trial of the 2015 assessment cycle (OECD, 2016). The construct of the paper-and-pencil and computer-based test was found to be consistent, the trend items were at most about 90% invariant in difficulty. At the item level, mode effect was found in both directions (easier or harder than the paper-and-pencil version), so the correction was made at the item level. The 6 related articles are of high measurement quality in terms of procedure and statistics used, but they are typically related to Germany and the 2012 and 2015 assessment cycles. The studies come to a similar conclusion, finding a non-systematic mode effect of 10–20 points on the achievement scores. The analysis of textual answers shows slightly longer and answers with more information in the case of computer-based tests.

For TIMSS 2019, one of the articles is the mode effect study itself (Fishbein et al., 2018). The mode effect was found to be negligible for the items, with a systematic difference of 7–14 points for the achievement points, showing that computer-based test is more difficult. On this basis, the constructs of the measures are the same, but in the case of trends, different corrections were applied for each year and measurement area. No systematic mode effect per country was found either, but at the same time, according to the analysis of the Dutch sub-sample, the common correction slightly overestimates the result (Robitzsch et al., 2020).

5.2. From linear to adaptive testing - the role of open-ended items

A prerequisite for adaptive testing is that the answers given by the students are immediately scored by the system, and ability is estimated on this basis. Approximately one third of the paper-and-pencil NABC tests items are open-ended items that require the work of a qualified coder. Open-ended items are not included to assess specific thinking operations or content areas more accurately, but to add to the diversity of the measure as a whole (Balázsi et al., 2014). Although open-ended items were gradually removed from the computer-based test, tasks that required short text responses (one word or number) and could be automatically scored remained, these item types were considered open-ended items in the present study.

The study was conducted on the 2017 NABC student data. After removing the data of unevaluated booklets and exempt students, in addition to the ability estimates calculated for the whole test, ability points were also calculated based on closed items only and assigned to the 7 + 1 NABC competency levels. The results were weighted according to the NABC methodology (N_6 = 86151, N_8 = 80886, N_{10} = 76550).

Based on the correlation analysis of the complete and closed item tests, reading and mathematical ability scores were correlated between 0.664 and 0.777, regardless of whether the given domain was measured with a complete test or only with closed items, indicating a moderate to strong relationship (Vargha, 2015) between the two domains, regardless of the use of open-enden items. Ability points calculated from closed questions and ability points calculated from the whole test show a correlation of more than 0.9 but less than 1 in both domains, which is considered very strong. On this basis, there is no difference between the constructs measured by the full test and the test consisting of closed items only.

By comparing the ability levels, we checked the discrepancy between the ability estimates. In both domains, the centralising effect of the use of closed items is evident, i.e. students at the lowest levels (below level 1 and level 1) are more likely to be moved up a level, while those at higher ability levels are moved down a level. 40% of the students at below level 1 and one third of the students at level 1 were placed in a higher ability level. In mathematics, the bias towards higher levels is already noticeable at level 5 and level 6, while in reading comprehension levels 6 and level 7 are affected. Only less than 1% of students had two levels of error.

5.3. Adaptive testing design – theoretical optimum

The result regarding the sample size and test length of adaptive tests is based on the idea that the objectives of a given assessment and the level to be achieved by its goodness indicators can be determined in advance. In the following, the starting point for the design is the test reliability indicator, i.e. the objective is to achieve a certain level of measurement error for each test administration, whether it is about the items (S_1) or the students (S_2). In the case of adaptive testing, the difficulty of the test can be planned, i.e. the probability of solving the next item (p), which is typically 50%, but other values are possible. Suppose that items and students are categorised by K ability levels, which is the precision of the measurement. The KR-20 and Wright formulae are used to measure reliability, so the test is based on the use of dichotomous items and the Rasch model.

Based on the result of the derivation, (1) the sample size (N) required for the measurement of the item can be calculated based on the expected accuracy of the item (S_1), the precision of the measurement (K), and the difficulty of the test (p), and (2) the expected accuracy of the ability estimation (S_2), the difficulty of the test (p), the sample size (N), and the precision of the measurement (K) can be used to calculate the expected length of the test (L).

(1) $$S_1 = \sqrt{\frac{N}{s_i(N-s_i)}} = \frac{1}{\sqrt{N}}\left[\frac{K}{\sqrt{p(K-p)}}\right],$$

(2) $$S_2 = \sqrt{\frac{1}{Lpq}}\sqrt{1 + \frac{\left(\frac{L}{L-1}\right)\left(\frac{N-L}{N}\right)^2 \ln^2\left(\frac{K-p}{p}\right)}{2{,}89}}.$$

In the case of items with a balanced solution (p = ½) typical of adaptive testing, the equations can be further refined:

(3) $$N = \left[\frac{1}{S_1^2}\frac{4K^2}{2K-1}\right],$$

(4) $$D = \frac{2{,}89}{(2K-1)},$$

$$A = \frac{2{,}89S_2^2}{4(2K-1)} = \frac{S_2^2}{4}D,$$

$$0 = L^3 - L^2(AN^2 + 2N) + L(N^2 + DN^2 + AN^2) - DN^2.$$

The required sample size and the expected length of the tests are increased by the higher precision and the higher number of ability levels, as well as the difficulty that differs from 50%. For example, in the case of NABC, where the precision of the measurement is 7 + 1 ability level, assuming a balanced test, in order to measure the

difficulty of an item with a standard error of 0.2, a sample of 400 people must be carried out, provided that they are distributed uniformly on the ability scale based on some ability estimate, i.e. not most respondents come from levels in the middle. In the case of the same assessment, with a standard error of 0.5 (100 points) in terms of the standard deviation of the ability estimate, it is expected that, with a sample of 100,000 student, each test will end in fewer than 58 items. With fewer than 5 ability levels, 44 items per test may be sufficient.

5.4. Comparison of possible adaptive strategies based on accuracy and reliability

In the final phase of my research, I conducted simulation tests to compare the effectiveness of some ability estimation and item selection methods. I ran a hybrid simulation in the sense that the item bank was made up by the well-functioning dichotomous items (625 items) of the 2008–2019 cycles and their parameters calculated according to the three-parameter model. The simulations thus also provide preliminary information as to whether the digitised versions of the items accumulated over the years are suitable for adaptive testing.

The theoretical ability of the students' was given by the fine division of the ability scale. Between 800 and 2200 points, I simulated two hundred measurements in each interval with a step interval of 50 points. The starting value of the simulated tests was the national average for grade 6, 1500 points. For the Bayesian ability estimation, I defined a prior distribution with a mean of 1500 points and a standard deviation of 200 points. The ability estimation procedures investigated were Maximum Likelihood (ML) (Lord, 1980), Bayes Modal (BM) (Birnbaum, 1969) and Expected a-Posteriori (EAP) (Bock & Mislevy, 1982). Among the item selection procedures, the maximum Fisher information (MFI) (Birnbaum, 1968), closest difficulty (bOpt) (Urry, 1970) and closest maximum information (thOpt) (Barrada et al., 2006) criteria were compared. Two possible measurement objectives and corresponding stopping criteria were used. In the first case, I examined the standard error of the estimation in addition to the 50 item tests. In the second case, the aim was to achieve uniform measurement accuracy, a standard error of 60 points, then the length of the tests could be compared.

Based on the results, maximum likelihood estimates performed best in terms of the average difference between theoretical ability and ability estimation. In the case of

fixed test length, the edge of the ability scale showed a difference of up to 100 points smaller than the Bayes estimates, in the case of fixed error the difference was only 50 points.

Among the item selection methods, Fisher's maximum information selection gave slightly better results than the nearest difficulty and nearest maximum information selection methods. This method produced shorter tests alongside a fixed error, and more accurate tests alongside the fixed test length in the range between 1,200 and 1,900 ability points than the current linear test.

To further examine the item bank consisting of the current items, a Monte Carlo simulation was used where an item bank of 300 items was simulated with similar but evenly distributed difficulty and an average slope that was 0.5 logit higher than expected for the NABC. The sample size was 90,000 and the combined stopping criterion was reaching 50 items or 60 points, in line with the previous study. In the simulation, the MFI selection method and BM and EAP estimation were used in addition to a two-parameter IRT model. To ensure the safety of the tests, an exposure rate of 20% was set.

According to the results, the tests ended after an average of 19–20 items, and there were hardly any tests longer than 30 items. The correlation between the theoretical and estimated ability scores is very high (r = 0.95). The standard deviation of the difference between the theoretical and the estimated ability, i.e. the standard error, is about 60 points (*RMSE* = 60.44), the average of the differences is -0.39 points, so there is no systematic difference. Regarding the exposure of the items, 44 items were administrated in the largest possible number, they were typically selected from the items with higher slope, while at the same time 65 items were not used at all in the simulation, they were typically less discriminating items.

6. Summary, limitations and directions for further research

In my research I aimed to investigate a possible direction for the development of the National Assessment of Basic Competencies based on digitisation, the methodological issues of computerised adaptive testing. To carry out preliminary tests preparing the successful transition from paper-and-pencil to computerised adaptive testing in the case of one of the Hungarian student assessment systems.

The transition from paper-and-pencil to computer-based assessment was completed during the research period, in 2022, but the summary of our results in relation

to the mode effect of international large-scale student assessments is still considered to address a void. The results of PISA and TIMSS assessments of mode effects are somewhat different: while TIMSS found a minimal mode effect, which varied in size by subject and grade, but showed that computer-based test was systematically more difficult, PISA found small deviations in different directions for a small number of items. Based on the results of the systematic literature review, there is no reason to expect a substantial mode effect in the case of NABC either, but small differences in difficulty parameters or ability points are possible, either by domain or in the case of individual items. An empirical study is recommended to investigate this issue.

Based on the comparison of ability points calculated from the whole test and only from closed-ended items, the tests measure the same phenomenon. The role of open-ended items in differentiating between students with the lowest and highest ability level, i.e. in identifying drop-outs and talents, may play a role, but this is not the aim of the NABC, partly because of the large number of individual errors. The higher scores of lower ability students may be due to guessing, which affects closed items, and willingness to answer, which may be lower in open-ended items. The lower scoring of higher ability students may be due to the greater difficulty of open-ended items or the complexity of this type of task. Since an item requiring a short text response, which is considered open-ended in a paper-and-pencil test, is considered an automatically scored item in a computer-based test, it is possible that an even smaller deviation can be expected. At the same time, further research into open-ended items and the development of automatically scored items of similar difficulty and appropriate content is warranted.

The theoretical derivation bridges the gap between the mathematical equations and the design of both simulations and real tests. As it typically assumes ideal conditions, it does not take into account, for example, the high uncertainty of the ability estimate at the beginning of the test, so its result can best be considered as a theoretical lower limit. At the same time, two and three parameter models can give better results than the use of the Rasch model, but their theoretical investigation is currently too complicated. If ability levels are used in the stopping criteria or item selection step, it may be recommended to drastically reduce the number of ability levels in order to shorten the test.

Based on the results concerning the mode effect and the study of open-ended items, the item- and student-level data collected in paper-and-pencil linear tests can probably be used well for the design of computerised or adaptive tests. Based on the simulation results, maximum likelihood estimation of ability and maximum Fischer

information item selection can lead to the most accurate and fastest tests. However, it can be assumed that in addition to the items developed for paper-and-pencil linear tests, which tend to focus on the middle part of the ability scale and measure a wider range of students, it is necessary to develop items that focus on the edges of the ability scale and/or have higher discrimination parameter. The validity of the simulations would be further improved by the use of real student responses or a more personalised choice of entry value that incorporates further information, so these are possible ways forward. The next step is also to map the other features of the item bank, which can be done outside the Educational Authority, as well as item development, which I recommend should continue to be done within the Authority for test security reasons.

NABC is committed to the development of the adaptive testing method from both a professional (Balázsi et al., 2021) and educational policy (Karkó, 2023) side, therefore the examination of the transition from the linear test to the adaptive testing is current and relevant. The information from the NABC is extensive due to its use, the examination of the topic has social utility. At the same time, adaptive testing, although it may result in shorter tests and more accurate ability estimation, can be expected to achieve only one of these two possible goals. That is why it would be necessary to define the purpose of the assessment more precisely before further developments. Shortening the test would reduce the burden on schools and students, certainly at middle ability levels, but the accuracy of individual results is not expected to be higher. Uniform accuracy of ability estimates would rather mean greater validity of individual results and the possibility to evaluate teachers, while also likely to result in somewhat shorter tests at intermediate levels.

References

Akhtar, H., Silfiasari, Vekety, B., & Kovacs, K. (2023). The Effect of Computerized Adaptive Testing on Motivation and Anxiety: A Systematic Review and Meta-Analysis. *Assessment*, *30*(5), 1379–1390. https://doi.org/10.1177/10731911221100995

Balázsi, I., Balkányi, P., Balogh, V. K., Gyapay, J., Ostorics, L., Palincsár, I., Rábainé Szabó, A., Suhajda, E., Szepesi, I., Szipőcsné Krolopp, J., & Velkey, K. (2021). *Folytonosság és változás az Országos kompetenciamérés szövegértés és matematika tartalmi kereteiben* (Vol. 1). Oktatási Hivatal.

https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/digitalis_orszmer/OKMtartalmikeret_Szovegertes_Matematika.pdf

Balázsi I., Balkányi P., Ostorics L., Palincsár I., Rábainé Szabó A., Szepesi I., Szipőcsné Krolopp J., & Vadász C. (2014). *Az Országos kompetenciamérés tartalmi keretei—Szövegértés, matematika, háttérkérdőívek*. Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2014/AzOKMtartalmikeretei.pdf

Barrada, J. R., Mazuela, P., & Olea, J. (2006). Maximum information stratification method for controlling item exposure in computerized adaptive testing. *Psicothema*, *18*(1), 156–159.

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, *6*(2), 258–276. https://doi.org/10.1016/0022-2496(69)90005-4

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Studies in Educational Evaluation*, *62*, 1–9. https://doi.org/10.1016/j.stueduc.2019.04.005

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika*, *12*(1), 1–16. https://doi.org/10.1007/BF02289289

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), Article 3. https://doi.org/10.1007/BF02310555

Csíkos C., & Vidákovich T. (2012). A matematikatudás alakulása az empirikus vizsgálatok tükrében. In Csapó B. (Ed.), *Mérlegen a magyar iskola* (pp. 83–130). Nemzeti Tankönyvkiadó.

DuToit, M. (Ed.). (2003). *IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact*. Scientific Software International.

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 Item Equivalence Study: Examining Mode Effects for Computer-Based Assessment

and Implications for Measuring Trends. *Large-Scale Assessments in Education, 6*. https://doi.org/10.1186/s40536-018-0064-z

Hood, C. (1991). A Public Management for All Seasons? *Public Administration, 69*(1), 3–19. https://doi.org/10.1111/j.1467-9299.1991.tb00779.x

Ito, K., & Segall, D. O. (2013). A Comparison of Four Methods for Obtaining Information Functions for Scores From Computerized Adaptive Tests With Normally Distributed Item Difficulties and Discriminations. *Journal of Computerized Adaptive Testing*, *1*(5).

Karkó, Á. (2023). Mindig minden változik: Az emberek, a társadalom és az oktatás. *Új Köznevelés*, *79*(7), 3–5.

Kehl D. (2012). Monte-Carlo-módszerek a statisztikában. *Statisztikai Szemle*, *90*(6), 521–543.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160. https://doi.org/10.1007/BF02288391

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Inc.

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized Adaptive and Multistage Testing with R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-69218-0

Magyar A. (2012). Számítógépes adaptív tesztelés. *Iskolakultúra*, *22*(6), Article 6.

Magyar A. (2014). Adaptív tesztek készítésének folyamata. *Iskolakultúra*, *14*(4), 26–35.

Magyar A., & Molnár G. (2015). A szóolvasási készség online mérésére kidolgozott adaptív és lineáris tesztrendszer összehasonlító hatékonyságvizsgálata. *Magyar Pedagógia*, *115*(4), Article 4. https://doi.org/10.17670/MPed.2015.4.403

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 Assessment Frameworks*. TIMSS & PIRLS.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2019). *PIRLS 2021 Assessment Frameworks*. TIMSS & PIRLS; Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/pirls2021/frameworks/

Mullis, I. V. S., Martin, M. O., & von Davier, M. (Eds.). (2021). *TIMSS 2023 Assessment Frameworks*. TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2023/frameworks/pdf/T23_Frameworks.pdf

Nagybányai-Nagy O. (2006). A pszichológiai tesztek reliabilitása. In Rózsa S., Nagybányai-Nagy O., & Oláh A. (Eds.), *A pszichológiai mérés alapjai: Elmélet, módszer és gyakorlati alkalmazás* (pp. 103–116). Bölcsész Konzorcium.

OECD. (2016). Annex A6 The PISA 2015 field trial mode-effect study. In *PISA 2015 Results (Volume I): Excellence and Equity in Education*. OECD. https://doi.org/10.1787/9789264266490-en

OECD. (2017). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD. https://doi.org/10.1787/9789264281820-en

OECD. (2019). *PISA 2018 Assessment and Analytical Framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

OECD. (2023). *PISA 2022 Assessment and Analytical Framework*. OECD Publishing. https://doi.org/10.1787/dfe0bf9c-en

Oktatási Hivatal. (2022, August 17). *A digitális országos mérések általános leírása*. https://www.oktatas.hu/kozneveles/meresek/digitalis_orszagos_meresek/altalanos_leiras

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., & Moher, D. (2021). Updating guidance for reporting systematic reviews: Development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology*, *134*, 103–112. https://doi.org/10.1016/j.jclinepi.2021.02.003

R Core Team. (2016). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Robitzsch, A., Luedtke, O., Goldhammer, F., Kroehne, U., & Koeller, O. (2020). Reanalysis of the German PISA Data: A Comparison of Different Approaches for Trend Estimation With a Particular Emphasis on Mode Effects. In *Frontiers in Psychology* (Vol. 11). FRONTIERS MEDIA SA. https://doi.org/10.3389/fpsyg.2020.00884

Rother, E. T. (2007). Systematic literature review X narrative review. *Acta Paulista de Enfermagem*, *20*, v–vi. https://doi.org/10.1590/S0103-21002007000200001

Şahin, A., & Weiss, D. J. (2015). Effects of Calibration Sample Size and Item Bank Size on Ability Estimation in Computerized Adaptive Testing. *Educational Sciences: Theory & Practice*, *15*(6), 1585–1595. https://doi.org/10.12738/estp.2015.6.0102

Sari, H. İ. (2020). Testing Multistage Testing Configurations: Post-Hoc vs. Hybrid Simulations. *International Journal of Psychology and Educational Studies*, *7*(1), 27–37. https://doi.org/10.17220/ijpes.2020.01.003

Thompson, N. A., & Weiss, D. A. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research, and Evaluation, 16*(1), 1–9. https://doi.org/10.7275/WQZT-9427

Urry, V. W. (1970). *A Monte Carlo investigation of logistic test models* [Nem publikált PhD-értekezés, Purdue University]. https://files.eric.ed.gov/fulltext/ED058317.pdf

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. https://doi.org/10.2458/v2i1.12351

Weiss, D. J., & Kingsbury, G. G. (1984). Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measurement*, *21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Wise, S. L. (2014). The Utility of Adaptive Testing in Addressing the Problem of Unmotivated Examinees. *Journal of Computerized Adaptive Testing*, *2*(1), 1–17.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Mesa Press.

Related publications

T. Kárász J., Nagybányai Nagy O., Széll K., & Takács S. (2022). Cronbach-alfa: Vele vagy nélküle? *Magyar Pszichológiai Szemle, 77*(1), 81–98. https://doi.org/10.1556/0016.2022.00004

T. Kárász J., & Széll K. (2023). Hogyan térnek el a papír-ceruza és számítógépes teszteredmények? - Szisztematikus szakirodalom áttekintés a PISA, TIMSS és PIRLS mérésekkel kapcsolatos tapasztalatokról. *Iskolakultúra*, *33*(3), 51–73.

T. Kárász, J., Széll, K., & Takács, S. (2023). Closed formula of test length required for adaptive testing with medium probability of solution. *Quality Assurance in Education*, *31*(4), 637–651. https://doi.org/10.1108/QAE-03-2023-0042

T. Kárász J., & Takács S. (2021). Adaptív tesztek minimális hosszának, hibájának, értékelési szintjének és a megoldók számának összefüggései—Általános megoldási aránnyal. *Alkalmazott Matematikai Lapok, 38*(1), 39–58. https://doi.org/10.37070/AML.2021.38.1.04

T. Kárász, J., & Takács, S. (2023). Use of open and closed items in automation of evaluation systems. *Alkalmazott Pszichológia, 25*(3), 33–54. https://doi.org/10.17627/ALKPSZICH.2023.3.33

Non-related publications

Péter P., Szivák J., Rapos N., & T. Kárász J. (2021). A pályakezdő tanárok tanulásának jellemzői. *Pedagógusképzés, 20*(3), 5–28. https://doi.org/10.37205/TEL-hun.2021.3.01

Rapos N., Szivák J., Tókos K., T. Kárász J., & Lénárd S. (2021). Az optimális tanulási környezet támogatásának intézményi lehetőségei vezetői és tanári nézőpontból. In Fehérvári A., Paksi B., & Széll K. (Szerk.), *Számít-e az iskola?* (pp. 87–104). Eötvös Loránd Tudományegyetem; MTMT. https://m2.mtmt.hu/api/publication/32187118

T. Kárász J. (2019). Hibabecslési eljárások véletlen jelenségek paramétereinek becslésére. *Psychologia Hungarica Caroliensis, 7*(2), 104–114. https://doi.org/10.12663/PSYHUNG.7.2019.2.7.

T. Kárász, J., Nagybányai-Nagy, O., Takács, N., & Takács, S. (2022). Egy felsőoktatási e-learning tananyagfejlesztés értékelése. *Educatio, 31*(2), 303–312. https://doi.org/10.1556/2063.31.2022.2.10

Takács, R., T. Kárász, J., Takács, S., Horváth, Z., & Oláh, A. (2021). Applying the Rasch model to analyze the effectiveness of education reform in order to decrease computer science students' dropout. *Humanities and Social Sciences Communications, 8*(1), 1–8. https://doi.org/10.1057/s41599-021-00725-w

Takács, R., T. Kárász, J., Takács, S., Horváth, Z., & Oláh, A. (2022a). Oktatási reform hatékonyságának vizsgálata – Tantárgyak nehézségi elemzése IRT-modell segítségével programtervező informatikus hallgatók körében. *Magyar Pszichológiai Szemle, 77*(2), 209–229. https://doi.org/10.1556/0016.2022.00014

Takács, R., T. Kárász, J., Takács, S., Horváth, Z., & Oláh, A. (2022b). Successful Steps in Higher Education to Stop Computer Science Students from Attrition. *Interchange* 53, 637–652. https://doi.org/10.1007/s10780-022-09476-2

Takács, R., Takács, S., T Kárász, J., Horváth, Z., & Oláh, A. (2021). Exploring Coping Strategies of Different Generations of Students Starting University. *Frontiers in Psychology, 12*. https://www.frontiersin.org/article/10.3389/fpsyg.2021.740569

Takács, R., Takács, S., T. Kárász, J., Oláh, A., & Horváth, Z. (2023). The impact of the first wave of COVID-19 on students' attainment, analysed by IRT modelling method. *Humanities and Social Sciences Communications, 10*(1), Article 1. https://doi.org/10.1057/s41599-023-01613-1

Takács, R., Takács, S., T. Kárász, J., Oláh, A., & Horváth, Z. (2024). Applying Q-methodology to investigate computer science teachers' preferences about students' skills and knowledge for obtaining a degree. *Humanities and Social Sciences Communications, 11*(1), 1–10. https://doi.org/10.1057/s41599-024-02794-z

Tókos K., Rapos N., Szivák J., Lénárd S., & T. Kárász J. (2020). Osztálytermi tanulási környezet vizsgálata. *Iskolakultúra, 30*(8), 41–61. https://doi.org/10.14232/ISKKULT.2020.8.41

Tókos, K., Takácsné Kárász, J., Rapos, N., Lénárd, S., & Szivák, J. (2023). Classroom learning environments and dropout prevention in Hungary. *European Journal of Education, 58*(4), 741–758. https://doi.org/10.1111/ejed.12591

Conference presentations on the topic

T. Kárász, J., & Széll, K. (2022a). MODE EFFECT AND ITEM EQUIVALENCE IN LARGE-SCALE INTERNATIONAL STUDENT ASSESSMENTS - A SYSTEMATIC LITERATURE REVIEW. In *EDULEARN22 Proceedings* (https://dx.doi.org/10.21125/edulearn.2022.1275; pp. 5399–5399). IATED. https://library.iated.org/view/TKARASZ2022MOD

T. Kárász, J., & Széll, K. (2022b). Nagymintás nemzetközi tanulói teljesítménymérések szisztematikus szakirodalmi áttekintése az elektronikus adatfelvétel szemszögéből. In D. Molnár & D. Molnár (Szerk.), *XXV. Tavaszi Szél Konferencia Absztraktkötet* (pp. 584). Doktoranduszok Országos Szövetsége (DOSZ).

T. Kárász, J., Széll, K., & Takács, S. (2022a). Adaptív tesztelés során szükséges teszthossz zárt formulája közepes megoldottsági valószínűség mellett. In D. Molnár & D. Molnár (Szerk.), *XXV. Tavaszi Szél Konferencia Absztraktkötet* (pp. 711). Doktoranduszok Országos Szövetsége (DOSZ).

T. Kárász, J., Széll, K., & Takács, S. (2022b). CLOSED FORMULA OF REQUIRED ITEM NUMBER FOR ADAPTIVE TESTING WITH MEDIUM PROBABILITY OF ITEM SOLUTION. In *EDULEARN22 Proceedings* (https://dx.doi.org/10.21125/edulearn.2022.1284; pp. 5432–5432). IATED. https://library.iated.org/view/TKARASZ2022CLO

T. Kárász, J., & Takács, S. (2019b). Nyílt és zárt végű itemek közötti kapcsolatok az Országos kompetenciamérés (2017-es) adatainak elemzése nyomán. In BME GTK (Szerk.), *I. Szakképzés és Oktatás: Ma – Holnap konferencia. Fejlődés és partnerség: Absztraktkötet* (pp. 122–123). BME Gazdaság- és Társadalomtudományi Kar.

T. Kárász J., & Takács S. (2021a). Adaptív tesztek minimális hosszának, hibájának, értékelési szintjének és a megoldók számának összefüggései – általános megoldás. In Molnár G. & Tóth E. (Szerk.), *A neveléstudomány válaszai a jövő kihívásaira* (pp. 525). MTA Pedagógiai Tudományos Bizottsága, SZTE Neveléstudományi Intézet. http://edu.u-szeged.hu/onk2021/download/ONK_CES_2021_Absztrakt_Kotet_-_Book_of_Abstarcts.pdf

T. Kárász, J., & Takács, S. (2021b). Adaptív tesztek minimális hosszának, hibájának, értékelési szintjének és a megoldók számának összefüggései – általános megoldási aránnyal. In J. Sass (Szerk.), *Út a reziliens jövő felé. A Magyar Pszichológiai Társaság XXIX. Országos Tudományos Nagygyűlése* (pp. 194–195). Magyar Pszichológiai Társaság. http://mptnagygyules.hu/wp-content/uploads/2021/08/MPT_kivonatkotet_2021_0825.pdf

Takácsné Kárász, J. (2023a). Adaptív teszt képességbecslési és feladatkiválasztási módszereinek összehasonlítása szimulációs módszerrel az Országos kompetenciamérés adatain. In L. Kasik & Z. Gál (Szerk.), *19. Pedagógiai Értékelési Konferencia Absztraktkötet* (pp. 66). Szegedi Tudományegyetem Neveléstudományi Doktori Iskola. https://www.edu.u-szeged.hu/pek2023/download/PEK_2023_CEA_2023_absztraktkotet.pdf

Takácsné Kárász, J. (2023b). Adaptív teszt képességbecslési és feladatkiválasztási módszereinek összehasonlítása szimulációs módszerrel az Országos kompetenciamérés adatain. In A. Bajzáth, K. Csányi, & J. Győri (Szerk.), *Elkötelezettség és rugalmasság: A neveléstudomány útjai az átalakuló világban* (pp. 423). MTA Pedagógiai Tudományos Bizottság, ELTE Pedagógiai és Pszichológiai Kar. https://onk2023.ppk.elte.hu/download/onk_absztraktok_VEGSO-10-26.pdf

Takácsné Kárász, J. (2023c). Adaptív tesztműködtetési eljárások összehasonlítása szimulációs módszerekkel az Országos kompetenciamérés adatain. In G. Kulcsár & V. D. Horváth (Szerk.), *Találkozás a változásban—Változások a találkozásban: A Magyar Pszichológiai Társaság XXX. Országos Tudományos Nagygyűlése—Kivonatkötet* (pp. 45–46). Magyar Pszichológiai Társaság. https://mpt.hu/wp-content/uploads/2023/09/Kivonatkotet_2023.pdf

Off-topic conference presentations

Gergely, B., T. Kárász, J., & Takács, S. (2019). Különböző mérési modellek az Országos kompetenciamérésben: Mi állhat a hibák hátterében? In BME GTK (Szerk.), *I. Szakképzés és Oktatás: Ma – Holnap konferencia. Fejlődés és partnerség: Absztraktkötet* (pp. 66–67). BME Gazdaság- és Társadalomtudományi Kar.

Gergely, B., T. Kárász, J., & Takács, S. (2021). Hol a hiba? Többdimenziós IRT modellek alkalmazása az Országos kompetenciamérésben. In J. Sass (Szerk.), *Út a reziliens jövő felé. A Magyar Pszichológiai Társaság XXIX. Országos Tudományos Nagygyűlése* (pp. 192–193). Magyar Pszichológiai Társaság. http://mptnagygyules.hu/wp-content/uploads/2021/08/MPT_kivonatkotet_2021_0825.pdf

Kispál, S., Gergely, B., T. Kárász, J., & Takács, S. (2021). Hátrányban vannak-e a halmozottan hátrányos helyzetűek az országos kompetenciamérésben? In J. Sass (Szerk.), *Út a reziliens jövő felé. A Magyar Pszichológiai Társaság XXIX. Országos Tudományos Nagygyűlése* (pp. 193). Magyar Pszichológiai Társaság. http://mptnagygyules.hu/wp-content/uploads/2021/08/MPT_kivonatkotet_2021_0825.pdf

Koltói, L., Harsányi, S. G., Nagybányai-Nagy, O., & Takácsné Kárász, J. (2019). Családi háttér és iskolai teljesítmény—Születni tudni kell? In E. Lippai (Szerk.), *Összetart a sokszínűség* (pp. 164). Magyar Pszichológiai Társaság.

Kövesdi, A., Kovács, D., & T. Kárász, J. (2021). Diszgráfia és írási nehézség előfordulása 6., 8., 10. Osztályos gyermekek körében 2012-2018 időszakban. In J. Sass (Szerk.), *Út a reziliens jövő felé. A Magyar Pszichológiai Társaság XXIX. Országos Tudományos Nagygyűlése* (pp. 153–154). Magyar Pszichológiai Társaság. http://mptnagygyules.hu/wp-content/uploads/2021/08/MPT_kivonatkotet_2021_0825.pdf

Kövesdi A., Kovács D., & Takácsné Kárász J. (2019). Az SNI-vel és BTM-mel diagnosztizált 6, 8, 10. Osztályos gyermekek iskolai teljesítménye. In Lippai E. (Szerk.), *Összetart a sokszínűség* (pp. 165). Magyar Pszichológiai Társaság.

Nádor, A., & T. Kárász, J. (2021). Teljesítményingadozás és annak háttértényezői az országos kompetenciamérés tükrében. In J. Sass (Szerk.), *Út a reziliens jövő felé. A Magyar Pszichológiai Társaság XXIX. Országos Tudományos Nagygyűlése* (pp. 154–155). Magyar Pszichológiai Társaság. http://mptnagygyules.hu/wp-content/uploads/2021/08/MPT_kivonatkotet_2021_0825.pdf

Nyitrai E., Takács N., & Takácsné Kárász J. (2019). Szülői bevonódás és iskolai teljesítmény. In Lippai E. (Szerk.), *Összetart a sokszínűség* (pp. 163). Magyar Pszichológiai Társaság.

Péter P., Szivák J., Rapos N., & T. Kárász J. (2022). A kezdő tanárok szakmai fejlődése és tanulása. In Steklács J. & Molnár-Kovács Z. (Szerk.), *21. Századi képességek, írásbeliség, esélyegyenlőség. Absztraktkötet* (pp. 65). MTA Pedagógiai Tudományos Bizottság – PTE BTK Neveléstudományi Intézet. https://konferencia.pte.hu/sites/konferencia.pte.hu/files/ONK_absztraktkotet_2022.pdf

Péter, P., Szivák, J., Rapos, N., & T. Kárász, J. (2023a). A kezdő tanárok szakmai fejlődése és tanulása az eredményes tanulás modellje mentén. In A. Bajzáth, K. Csányi, & J. Győri (Szerk.), *Elkötelezettség és rugalmasság: A neveléstudomány útjai az átalakuló világban* (pp. 381). MTA Pedagógiai Tudományos Bizottság, ELTE Pedagógiai és Pszichológiai Kar. https://onk2023.ppk.elte.hu/download/onk_absztraktok_VEGSO-10-26.pdf

Péter, P., Szivák, J., Rapos, N., & T. Kárász, J. (2023b). Professional Development And Learning Of Novice Teachers. In E. P. Chaw, F. N. Barcın, L. A. Erdei, A. O.

Pongor-Juhász, & E. Kopp (Szerk.), *ATEE Annual Conference 2023: Teacher Education on the Move* (pp. 273–275). Association for Teacher Education in Europe (ATEE). https://ateeannual2023.elte.hu/wp-content/uploads/2023/09/ATEE%20Annual%20Conference%202023_Book%20of%20abstracts.pdf

Smohai M., Simon G., & Takácsné Kárász J. (2019). Az Országos kompetenciamérés során felvett szabadidős sporttevékenységre irányuló adatok elemzése. In Lippai E. (Szerk.), *Összetart a sokszínűség* (pp. 164). Magyar Pszichológiai Társaság.

T. Kárász J., Nagybányai-Nagy O., Takács N., & Takács S. (2021). Egy elsőéves egyetemi gyakorlat átalakítása a távolléti oktatás igényeinek és lehetőségeinek fényében. In Buda A. & Kiss E. (Szerk.), *Interdiszciplináris Pedagógia a bizonytalanság korában* (pp. 55). Debreceni Egyetem Nevelés- és Művelődéstudományi Intézet.

T. Kárász, J., & Takács, S. (2019a). Kevert mérési területek pilot vizsgálata meredekségi és nehézségi paraméterek elemzésével az Országos kompetenciamérés 2017-es adatai alapján. In BME GTK (Szerk.), *I. Szakképzés és Oktatás: Ma – Holnap konferencia. Fejlődés és partnerség: Absztraktkötet* (pp. 123). BME Gazdaság- és Társadalomtudományi Kar.

Tókos, K., Rapos, N., Szivák, J., Lénárd, S., & T. Kárász, J. (2021). An Examination of Classroom Learning Environments. In *(Re)imagining & Remaking Teacher Education* (pp. 223–224). Association for Teacher Education in Europe (ATEE). https://drive.google.com/file/d/1QdMROSsN5gsBXdIDPiAd7xACYqiVhd7i/view

Tókos K., T. Kárász J., Rapos N., Szivák J., & Lénárd S. (2021). Az osztálytermi tanulási környezet vizsgálata. In Molnár G. & Tóth E. (Szerk.), *A neveléstudomány válaszai a jövő kihívásaira* (pp. 437). MTA Pedagógiai Tudományos Bizottsága, SZTE Neveléstudományi Intézet. http://edu.u-szeged.hu/onk2021/download/ONK_CES_2021_Absztrakt_Kotet_-_Book_of_Abstarcts.pdf

Appendices

Appendix 1.

A systematic literature review on the mode effect

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 Item Equivalence Study: Examining Mode Effects for Computer-Based Assessment and Implications for Measuring Trends. L*arge-Scale Assessments in Education, 6*. https://doi.org/10.1186/s40536-018-0064-z

Hamhuis, E., Glas, C., & Meelissen, M. (2020). Tablet assessment in primary education: Are there performance differences between TIMSS' paper-and-pencil test and tablet test among Dutch grade-four students? *British Journal of Educational Technology, 51*(6), 2340–2358. https://doi.org/10.1111/bjet.12914

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice, 23*(4), 495–518. https://doi.org/10.1080/0969594X.2016.1147420

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education, 44*(4), 476–493. https://doi.org/10.1080/03054985.2018.1430025

Kroehne, U., Buerger, S., Hahnel, C., & Goldhammer, F. (2019). Construct Equivalence of PISA Reading Comprehension Measured With Paper-Based and Computer-Based Assessments. *Educational Measurement: Issues & Practice, 38*(3), 97–111. https://doi.org/10.1111/emip.12280

Robitzsch, A., Luedtke, O., Goldhammer, F., Kroehne, U., & Koeller, O. (2020). Reanalysis of the German PISA Data: A Comparison of Different Approaches for Trend Estimation With a Particular Emphasis on Mode Effects. *Frontiers In Psychology, 11*. FRONTIERS MEDIA SA. https://doi.org/10.3389/fpsyg.2020.00884

Zehner, F., DIPF, Kroehne, U., Hahnel, C., & Goldhammer, F. (2020). PISA reading: Mode effects unveiled in short text responses. *Psychological Test and Assessment Modeling, 62*(1), 85–105. Publicly Available Content Database.

Zehner, F., Goldhammer, F., Lubaway, E., & Sälzer, C. (2019). Unattended consequences: How text responses alter alongside PISA's mode change from 2012 to 2015. *Education Inquiry, 10*(1), 34–55. https://doi.org/10.1080/20004508.2018.1518080