SOCIETY. INTEGRATION. EDUCATION
Proceedings of the International Scientific Conference. Volume I, May 26th, 2023. 613-622

DEVELOPMENT OF COMPUTER-BASED DIAGNOSTIC ASSESSMENT SYSTEM: CASE STUDY OF EQUIVALENCE OF PAPER-AND-PENCIL AND COMPUTER-BASED TESTING

Ģirts Burgmanis
University of Latvia, Latvia

Marta Mikite
University of Latvia, Latvia

Ilze France
University of Latvia, Latvia

Dace Namsone
University of Latvia, Latvia

Abstract. *In the last two decades computer-based assessment has become an important part in support of teaching and learning. It is seen as a solution to implement assessment for learning in school and provide immediate feedback on students' performance in real-time. Research literature on computer-based assessment suggests that every measurement instrument developer before implementation of a test has to provide evidence that computer-based and paper-based versions are equivalent and provide consistent measures. There is a risk that properties of computer-based assessment including unfamiliarity with the system and proficiency level of digital skills can seriously affect students' performance. This paper focuses on computer-based diagnostic assessment system designed to support numeracy and literacy teaching and learning. The aim of this study is to confirm that literacy and numeracy learning measurement instruments elaborated in diagnostic assessment system provide consistent results as paper-based versions of both instruments. Data were collected administering four tests. Two of the assessments were computer-based literacy and numeracy diagnostic assessments and two were paper-based versions. By analyzing both versions of assessments using various statistical techniques we explore differences in students' performance. Our results showed that at this development phase of the computer-based diagnostic assessment system the students who completed computer-based test versions showed similar or better performance than their counterparts who completed paper-based versions.*
Keywords: *assessment, computer-based assessment, diagnostic, literacy, numeracy, paper-based assessment*

https://doi.org/10.17770/sie2023vol1.7096

Introduction

In today's digital age, technologies are becoming more important in both everyday school lessons and assessment. A digital diagnostic tool was created that measures students' numeracy and literacy in different subjects. Digitized assessment has many advantages as well as several significant risks. There is no unequivocal answer in previous studies about in which test mode students demonstrate higher performance, or how significant is the difference (Gallagher, Bridgeman, & Cahalan, 2002, McDonald, 2002). The research question investigated in this study is whether literacy and numeracy learning measurement instruments elaborated in computer-based diagnostic assessment system provide consistent results as paper-based versions of both instruments.

Literature review

The main advantages of computer-based tests (CBT) are that they are location independent, they provide immediate grading, can offer dynamic and individualized assessment, and can also help students with disabilities (Way, Davis, Keng, & Strain-Seymour, 2015). CBT also have some limitations due to external factors such as system problems, loss of electricity or internet. There are also limitations in task design and there is a risk that test mode affects student performance (Smolinsky, Marx, Olafsson, & Ma, 2020; McClelland & Cuevas 2020; McDonald, 2002). It must be assured that they reflect on a student's content proficiency, not on computer proficiency (including typing versus handwriting), because it affects the construct being measured and the interpretation of the obtained results can be misleading (Puhan, Boughton, & Kim, 2007; Gallagher et al., 2002).

Bennett describes the three stages of tehnology-based assessment (TBA) development (Bennett, 1998; Bennett, 2015). First-generation TBA is mostly related to the development of an appropriate infrastructure which includes a huge investment in computer hardware. Equally essential in technology staff to install and troubleshoot testing software, and training teachers to administer and manage online exams, including on how to deal with technology failure. (Bennett, 2015; Drasgow, Luecht, & Bennett, 2006). The tests themselves at this level are simple and a very similar to paper-pencil tests. In this generation, adaptive tests are being developed, which means that students' answers influence next test items. Adaptive tests can be designed to improve both measurement quality and measurement efficiency (Weiss, 1982). In second generation CBT are used to measure key competences, like information literacy (Bennett, 1998; Bennett, 2015). The need to measure new constructs leads to a change in the design of items from traditional to more interactive. Second-generation tests include qualitative (but incremental) change and efficiency improvement become the driving goals (Bennett, 2015).

The third generation of TBA can be characterized by three key elements: (1) these assessments serve both institutional and individual-learning purposes, (2) they are designed from cognitive principles and theory-based domain models, and (3) the assessments use complex simulations and other interactive performance tasks. In the third generation, the differences between instruction and assessment becomes blurred as continuous assessment occurs throughout the learning process (McDonald, 2002). What was, at first, an evolution driven primarily by technology becomes driven by content (Bennett, 2015).

The long evolution of technology-based assessment has led to a wide range of item types. In addition to multiple-choice and essay type items, there are increasingly used sophisticated TBA solutions such as game-based assessment and simulation-based assessment. In research (Popp, Tuzinski, & Fetzer, 2015.) authors explore issues that test developer should consider choosing between three most common test formats: text, video, and animation. There are four areas to consider when developing new simulation in framework developed by them: (1) psychometric, (2) applied, (3) contextual and (4) logistical. Others (Parshall & Harmes, 2008) proposes that the following aspects be considered when introducing innovations: (1) assessment structure, (2) response action, (3) media inclusion, (4) interactivity, (5) complexity, (6) fidelity, and (7) scoring method. Each element relates to important decisions that test developers must make when designing innovative items and their associated interfaces Parshall & Harmes, 2008; Popp et al., 2015). In general, the most promising benefit to any type of innovation is the potential to improve the measurement of the underlying construct (Parshall & Harmes, 2008).

During this transitional stage when technology is becoming commonplace in schools, but state, districts and schools have not made a complete switch to computerized assessments, CBT and paper-based test (PPT) co-exist. Comparability in their assessment of student performance is crucial. To be able to say with certainty that a CBT really measures what it is intended to, an analysis is needed that compares the results to PBT (Smolinsky et al., 2020). This is essential for validating the diagnostic results and for creating further tests. Previous studies have been conducted to determine comparability between the test modalities and some indicate that CBT and PPT scores are comparable, while others indicate a performance advantage for either CBT or PPT (McClelland & Cuevas, 2020).

Has been identified (McDonald, 2002) two fundamental types of equivalence which need to be examined: (1) score equivalence and (2) eligibility to the underlying construct. Equivalence in scoring can be observed by an examination of central tendency and score distributions, showing that score of 100 on a PBT test may be equivalent to a score of 90 on an apparently identical CBT (McDonald, 2002). Unlike differences in scoring which can be resolved relatively easily, differences in the constructs being measured cannot be resolved by

statistical methods, although they can be identified by analyzing statistical parameters like Rank ordering of test takers, reliability coefficients and the factor structure of the tests (McDonald, 2002).

Statistical investigations of equivalence have largely ignored the fact that in presenting a test on computer, a qualitatively different testing experience is created (McDonald, 2002). For example, "Given that writing is a cognitively complex and socially situated activity, it is clearly impossible to achieve complete equivalence between the two conditions." (Chan, Bax, & Weir, 2018). There is evidence that lower - performing individuals will be disadvantaged when carrying out computer-based assessment, in opposite of high-attaining students who performed better in CBT than PBT (Clariana & Wallace, 2002; Noyes, Garland & Robbins, 2004). These factors influencing the comparability of CBT and PBT are subject to constant and significant change as technology availability and usage patterns change. It is therefore particularly important to check the validity of CBT. When comparing CBT and PBT, it should not be forgotten that there are also physical differences between CBTs and PBTs, e.g. PBTs display all questions on a sheet, while CBTs often display questions one by one. CBT also has limited possibility to go back to previous questions or skip questions, while it can be done freely in PBTs (McDonald, 2002).

Computer-based diagnostic assessment system

First version of computer-based diagnostic assessment system which we test in this study was built in 2021 by Interdisciplinary Centre for Educational Innovation of University of Latvia and 'Izglītbas sistēmas' owner of largest digital school management system 'e-klase.lv' in Latvia. The system is designed for diagnostic purposes to assess students' literacy and numeracy learning in grade 4, 7 and 10.

In present form system is a technology-based, learning-centred and integrated assessment system consisting of four modules: (1) test editing module, (2) online test delivery module, (3) scoring module, (4) feedback module.

The tests can be ran by students using computers with equipped with an internet browser, keyboard, mouse and screen. To access test students need to login in system using their 'e-klase.lv' user login details. Each test can be administered by teachers who can choose the day and time when students have access to test and complete it. The system is designed for both automated and human scoring.

The items were written and saved in open source software GeoGebra applet and linked with system's test editing and item delivery module using applet ID generated by GeoGebra. It means that students respond to items in GeoGebra environment elaborated in system. Thus, the focus of the present study is to examine does choice of GeoGebra environment for item delivery module have

any effect on students' performance and may affect assessment validity and reliability.

Research Methodology

Participants

In April 2022, PBT and CBT literacy and numeracy tests were completed in nine secondary schools in Latvia. To ensure representative study sample all schools for both study samples were selected based on schools' overall performance in previous year's (2021) national level assessment in grade 6th in Latvian language and mathematics. The sample of study to test equivalence of PBT and CBT *literacy* tests included 766 students from grade 7. 519 students from five secondary schools each completed diagnostic assessment consisting of three CBT tests examining literacy skills in literature, history and science contexts. The same tests in PBT version completed 247 7th grade students from four other schools.

The sample of study to test equivalence of PBT and CBT numeracy tests included 712 students from grade 7. 505 students from 5 secondary schools each completed three CBT tests examining numeracy skills in mathematics (2 tests) and science contexts. 236 7th grade students from four other schools completed the same tests in PBT version.

Procedure

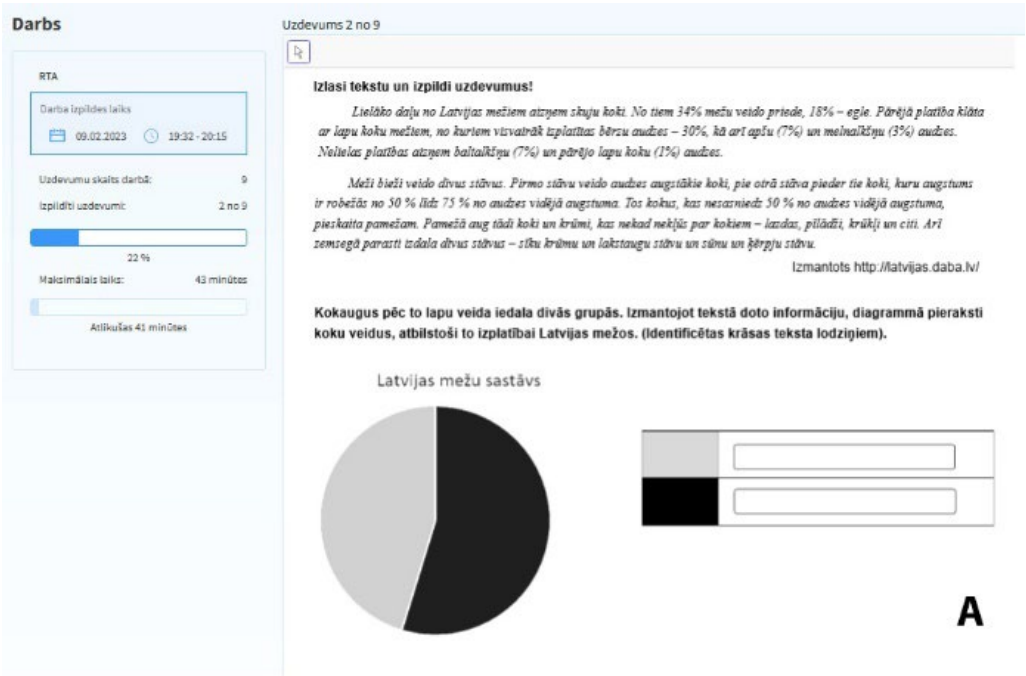
At the beginning of CBT tests, students were provided with instructions about the use of the system and allowed to familiarize its functionality completing several training items. PBT tests took place in the schools' ICT labs using the available school infrastructure. Both PBT and CBT testing sessions were supervised by teachers. Teachers can choose how to administer tests, i.e., one test per day or all tests in one day. Two experts separately based on previously developed and evaluated marking scheme scored all students' responses on test items. If scores for student's response on item for both experts differed then they agreed on the score.

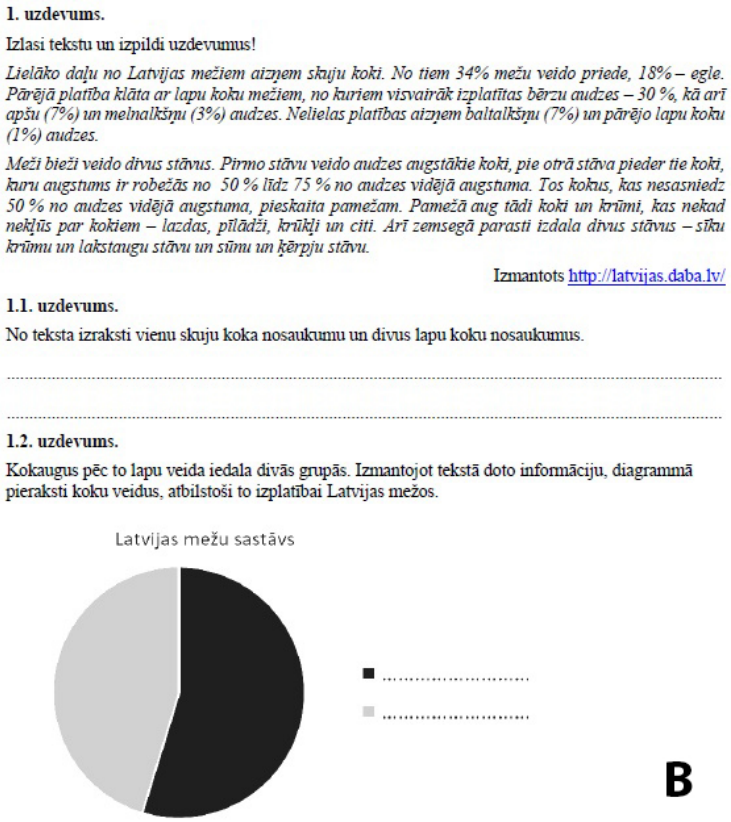
Instruments

Each of PBT and CBT tests were prepared to be completed in 45 minutes. However, if some of students cannot complete the test in time teacher allowed to complete it. Literacy knowledge and skill diagnostic assessment consisted of 45 items in total, i.e., literature test consisted of 22 items, history test included 15 items and science test – 8 items. Numeracy learning diagnostic assessment consisted of 37 items. First test of mathematics consisted of 13 and second test 17 items as well as science test of 7 items. Each test consisted of multiple-choice items and constructed response items.

Analysis

To examine research question we compared students' performance in CBT version items and in PBT version items. Figure 1 shows the same item in different modes. In our analysis we selected 11 items from which were 6 multiple-choice items and 5 constructed response items. Three of multiple-choice items were from literacy diagnostic assessment tests and three from numeracy assessment tests. Each item represented one test with particular context (literacy – history, literature, science; numeracy – mathematics: numbers, mathematics: ratios and relationships, science). Items were selected based on the number of students' responses. Items with higher number of responses were included in test sample. The similar approach was used to select constructed response items. However, in case of literacy only two items were selected, i.e., from history and science contexts.




1. uzdevums.

Izlasi tekstu un izpildi uzdevumus!

Lielāko daļu no Latvijas mežiem aizņem skuju koki. No tiem 34% mežu veido priede, 18% – egle. Pārējā platība klāta ar lapu koku mežiem, no kuriem visvairāk izplatītas bērzu audzes – 30 %, kā arī apšu (7%) un melnalkšņu (3%) audzes. Nelielas platības aizņem baltalkšņu (7%) un pārējo lapu koku (1%) audzes.

Meži bieži veido divus stāvus. Pirmo stāvu veido audzes augstākie koki, pie otrā stāva pieder tie koki, kuru augstums ir robežās no 50 % līdz 75 % no audzes vidējā augstuma. Tos kokus, kas nesasniedz 50 % no audzes vidējā augstuma, pieskaita pamežam. Pamežā aug tādi koki un krūmi, kas nekad nekļūs par kokiem – lazdas, pīlādži, krūkļi un citi. Arī zemsegā parasti izdala divus stāvus – sīku krūmu un lakstaugu stāvu un stūmu un ķērpju stāvu.

Izmantots http://latvijas.daba.lv/

1.1. uzdevums.

No teksta izraksti vienu skuju koka nosaukumu un divus lapu koku nosaukumus.

.....

.....

1.2. uzdevums.

Kokaugus pēc to lapu veida iedala divās grupās. Izmantojot tekstā doto informāciju, diagrammā pieraksti koku veidus, atbilstoši to izplatībai Latvijas mežos.

Latvijas mežu sastāvs

.....

.....

B

Figure 1 ***Example of literacy diagnostic assessment science test item (the same item for both versions: A – CBT version item, B – PBT version item)*** *(created by authors)*

To compare students' performance in CBT version and PBT version we align students' scores in PBT version constructed response items with CBT version of the same items. In CBT version, multiple-choice items as well as constructed response items were scored 0 or 1 where 1 were scored for constructed response items only when student responses were fully correct. At the same time in PBT version constructed response items were scored using marking scheme including other scores than 0 and 1. Thus, we rescored PBT version constructed items similarly as in CBT version.

Statistical analysis was performed for each assessment separately.

Research results

The equivalence between PBT and CBT versions developed for our diagnostic assessment system was confirmed by independent samples t-test. Table 1 illustrate the differences in students' performance in PBT and CBT versions of literacy diagnostic assessment. Foremost, we tested hypothesis stating that students' performance differs when they respond to multiple-choice items in PBT version and the same items in CBT version.

From Table 1, it can be seen that students' performance was significantly higher for PBT version only when they responded to multiple-choice item in

literature context. In other two contexts, there were no significant differences in students' performance between PBT and CBT version multiple-choice items.

Table 1 ***T-test results: students' performance in CBT and PBT versions of literacy diagnostic assessment items*** *(created by authors)*

	CBT		**PBT**		
	M	SD	M	SD	p
MC L Item	0,19	0,39	0,30	0,46	0,001**
MC S Item	0,60	0,49	0,66	0,47	0,204
MC H Item	0,74	0,44	0,75	0,44	0,897
CR H Item	0,54	0,50	0,38	0,49	0,001**
CR S Item	0,17	0,38	0,18	0,39	0,767

*Note: **significant at level p<0,05. Type of item – MC: multiple-choice item, CR: constructed response item; Context of test – L: literature, S: science, H: history.*

For numeracy, assessment students' performance is higher only for multiple-choice item from PBT version of mathematics test on ratios and relationships context (see Table 2). Students who completed PBT version of item show higher performance than students who responded the same item in CBT version. Furthermore, results show that students' performance who completed CBT versions of multiple-choice items from other two tests are similar to those who completed PBT version or higher (science).

Table 2 ***T-test results: students' performance in CBT and PBT versions of numeracy diagnostic assessment items*** *(created by authors)*

	CBT		PBT		
	M	SD	M	SD	p
MC S Item	0,17	0,38	0,08	0,28	0,008**
MC MT1 Item	0,92	0,27	0,91	0,28	0,656
MC MT2 Item	0,59	0,49	0,73	0,44	0,004**
CR S Item	0,57	0,49	0,65	0,48	0,109
CR MT1 Item	0,04	0,19	0,12	0,33	0,011**
CR MT2 Item	0,14	0,35	0,12	0,32	0,632

*Note: **significant at level p<0,05. Type of item – MC: multiple-choice item, CR: constructed response item; Context of test – MT1: mathematics/numbers, MT2: mathematics/ratios and relationships S: science.*

Next, we tested the effect of test version on student performance same in case of constructed response items. Table 1 reveal that students who completed literacy diagnostic assessment in CBT version show higher or similar performance for constructed response items than students who completed the same items in PBT version. For numeracy diagnostic assessment only in mathematics test on numbers context constructed response item were answered more correctly in PBT version than CBT version. At the same time students'

performance were similar for PBT and CBT versions for constructed response items in both other tests.

Conclusions

Development of computer-based assessment is important to support learning and teaching and provide immediate feedback on students' performance in real-time. However, there is still a question does CBT version of assessment can provide consistent evidence with PBT version of assessment. This study focused on computer-based assessment system developed by Interdisciplinary Centre for Educational Innovation of University of Latvia and 'Izglītbas sistēmas' to test students learning of literacy and numeracy knowledge and skills. In this study, we examined and compared CBT and PBT versions of diagnostic assessment of both skills. The results of our study will be applied in further improvement of system. Our study revealed that in most cases CBT versions of multiple-choice items and constructed response items provide similar results as PBT versions. Moreover, this pattern did not differ between literacy and numeracy diagnostic assessments. In some cases, CBT versions of items were responded better than the same PBT versions. However, the study also showed that some items should be purified in CBT version to provide more consistent results with PBT version. Thus, we can confirm that GeoGebra environment elaborated in item delivery module have little negative effect on students' performance and in some cases can even help to get better results.

Finally, we can conclude that existing CBT versions of both diagnostic assessments can be used as example for development of similar diagnostic instruments to measure literacy and numeracy skills in other grades in future.

Acknowledgements

The research was promoted with support of the European Regional Development Fund's project 'IT-based support system prototype for providing feedback and improve student performance in literacy and numeracy acquisition', Project No. 1.1.1.1/19/A/076.

References

Bennett, R. E. (1998). *Reinventing Assessment. Speculations on the Future of Large-Scale Educational Testing. A Policy Information Perspective.* Princeton, NJ: Educational Testing Service, Policy Information Center

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, *39*(1)*,* 370-407.

Chan, S., Bax, S., & Weir, C. (2018). Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test. *Assessing Writing*, *36*, 32-48.

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

Drasgow, F., Luecht, R. M., & Bennett, R.E. (2006). Technology and testing. *Educational measurement, 4,* 471-515.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. *Journal of Educational Measurement, 39*(2)*,* 133-147.

McClelland, T., & Cuevas, J. A. (2020). A comparison of computer based testing and paper and pencil testing in mathematics assessment. *The Online Journal of New Horizons in Education, 10*(2), 78-89.

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education, 39*(3), 299-312.

Noyes, J., Garland, K., & Robbins, L. (2004). Paper-based versus computer-based assessment: is workload another test mode effect? *British Journal of Educational Technology*, *35*(1), 111-113.

Parshall, C. G., & Harmes, J. C. (2008). The design of innovative item types: Targeting constructs, selecting innovations, and refining prototypes. *CLEAR Exam Review, 19*(2), 18-25.

Popp, E. C., Tuzinski, K., & Fetzer, M. (2015). Actor or Avatar?: Considerations in Selecting Appropriate Formats for Assessment Content. In F. Drasgow (Eds.), *Technology and testing: Improving educational and psychological measurement.* (pp. 79–103). Abingdon, UK: Routledge.

Puhan, G., Boughton, K., & Kim, S. (2007). Examining Differences in Examinee Performance in Paper and Pencil and Computerized Testing. *Journal of Technology, Learning, and Assessment*, *6*(3), 1-21.

Smolinsky, L., Marx, B. D., Olafsson, G., & Ma, Y. A. (2020). Computer-based and paper-and-pencil tests: A study in calculus for STEM majors. *Journal of Educational Computing Research, 58*(7), 1256-1278.

Way, W. D., Davis, L. L., Keng, L., & Strain-Seymour, E. (2015). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Eds.), *Technology in testing: Improving educational and psychological measurement (Vol. 2)*. (pp. 260-284). Abingdon, UK: Routledge.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied psychological measurement, 6*(4), 473-492.