Macrothink Institute™

Journal of Studies in Education
ISSN 2162-6952
2020, Vol. 10, No. 1

The Developmental Tendency in Computer-based Assessment

Saif Husam Mohammed (Corresponding author)

Doctoral School of Education, University of Szeged

Szeged, Aradi vértanúk tere 1, 6720 Hungary

E-mail: mohammed.saif.husam@edu.u-szeged.hu

Saleh Ahmad Alrababah

Doctoral School of Education, University of Szeged

Szeged, Aradi vértanúk tere 1, 6720 Hungary

E-mail: alrababah.saleh.ahmed@edu.u-szeged.hu

Received: Dec. 20, 2019 Accepted: Feb. 20, 2020 Published: February 25, 2020

doi:10.5296/jse.v10i1.16421 URL: https://doi.org/10.5296/jse.v10i1.16421

Abstract

This study will examine the primary characteristics of traditional assessment while outlining the benefits, limitations, and possibilities of computer-based assessment (CBA) concerning education. It will also assess the emerging technology regarding item writing in terms of achievement tests, along with the diverse approaches for item development. Further, numerous studies have focused on CBA's benefits for the classic methods of assessment. As noted by Molnár and Csapó (2018), information and communication technologies, computers in particular, significantly affect the development of educational examination from the quantitative as well as qualitative perspectives (Molnár & Csapó, 2018). Although CBA offers efficient examinations compared to traditional methods such as PP or face-to-face assessment, numerous CBA stages can be detected in PP's transition to third-generation CBA concerning the educational context. While the first generation CBA made less use of technology with its items, primarily multiple-choice, as well as tests being fixed and similar to PP tests and items, the second generation CBA tasks involved multimedia elements that made adaptive testing feasible, Moreover, the third generation CBA tasks ensured that complicated constructs could also be assessed such as the 21st-century skills through simulation, interaction, dynamically changing items, as well as cooperation (see Csapó &

Molnár, 2017). Apart from the item and test development options, technology opens a new arena through storing and assessing contextual data, known as educational data mining, learning analytics, or logfile analyses, which indicates a diverse analytical form. Considering the numerous advantages, the crucial assessments in the coming future may be implemented through a technological environment.

Keywords: paper-based testing, computer-based assessment, contextual data, logfile analyses

1. Introduction

Over the last twenty years, the use of technology-based assessment in special education has undergone a radical change (Dikusar, 2018). Current applications are employed to support numerous functions and features, whereas earlier uses focused simply on managing test scores. The features incorporated today include self-administration functions, decision-making according to predetermined criteria, software-managed item presentation, algorithm-based response evaluation, prescription founded on expertise, and straightforward links between instructional amendments and assessment (Roschelle, Pea, Hoadley, Gordin, & Means, 2000).

Such assessments typically use systems and software to assess individual students within the educational environment, thereby combining both electronic renderings of established metrics new computer-based evaluation (McCain, 1995). One instance of evaluation employing technology is the use of video-based computer-aided testing as a means to ascertain students' language preferences so that ensuing testing can automatically be conducted in accordance. Another example employs video clips from popular films to present moral dilemmas for problem-solving exercises. Still, another instance requires students to observed short recordings and then to respond to questions using touch screens (Kraslawski & Turunen, 2013).

Assessment in conventional education systems was somewhat teacher-focused, but recently the model has moved to a more participatory approach in which student feedback is not only more individualized, involved, and relevant but also allows teachers to comprehend students' learning graphs (Alliance, 2017).

Hence, technology-based assessment can add authority and bear to testing methods. This is important due to the recognition that existing assessment processes have led to students being misallocated within the educational system. Student misplacement has profound consequences for both students and staff. For example, students might withdraw from education and abandon it altogether. For teachers, misallocation can result in the presence in their classroom of students with learning difficulties with which they have not been trained to assist (McCain, 1995).

2. Brief History of Computer-Based Testing

The College Board's ACCUPLACER® testing program was one of the original major computerized-adaptive testing programmes to be implemented. Launched in 1985, this programme was composed of four tests: (1) Reading Comprehension, (2) Sentence Skills, (3) Arithmetic, and (4) Elementary Algebra. By examining these four areas, future college applicants were able to be placed appropriately within English and Mathematics courses. Therefore, initially, the test had lower stakes. Novell Corporation's certified network engineer (CNE) examination was the first high stake computerized adaptive testing (CAT) (College Board, 2017). Luecht and Sireci (2011) reported that Drake Prometric testing centers launched CNE onto an online platform in 1990, which later progressed to online CAT in 1991. Education Testing Service's (ETS) Graduate Record Examination (GRE) followed the CNE.

According to Mills and Stocking (1996), from 1992, this was implemented across the USA as CAT within the Sylvan testing centers. In 1994, nursing candidates were impacted by two NCLEX examinations under the structure of a CAT, which occurred within commercial testing centers. As reported by Sands, Waters, and McBride, (1997), the Military Entrance Processing Stations were the focus for the CAT version of the Armed Services Vocational Aptitude Battery (ASVAB).

To drive the further progression of this initiative, a 1997 CAT edition of GMAT was introduced by the Graduate Management Admission Council. In 1997, the Architect Registration Examination (ARE) was introduced. This test presents collaborative and computer-driven architectural issues positioned within a personalized graphical platform. Having launched this architect's testing scale, the United States Medical Licensing Examination (USMLE) adapted and progressed to CBT during 1999 (Ejim, 2018). According to Clyman, Melnick, and Clauser (1995), this particular test consisted of advanced and interactive technology with computer-generated patient-management recreations. With such on-going progress within this field, the 2004 Uniform CPA examination incorporated interactive accounting simulations (DeVore, 2002). The original computer-adaptive multi-stage testing frameworks for major project work were also implemented (Breithaupt, Ariel, & Hare, 2009; R. Luecht, Brumfield, & Breithaupt, 2006; Melican, Breithaupt, & Zhang, 2009). A number of CBTS were launched within the licensure and certification environment and are demonstrated by these CBT programmes.

The early generation of computerized adaptive testing (CAT) had a significant impact whereby the early tests founded on computers benefited from more efficient computers and technologies. Rather than continue with handwritten testing, computer-generated tests encourage greater precision and more concise examinations. Similar to the old-school hand-written testing forms, the use of multiple fixed test forms was implemented within other examinations, including the Physical Therapist licensure exam. In recent times, the computer-adaptive multistage testing (ca-MST) has become widely recognized by institutions looking for an amalgamated version of testing methods. This unites the advantages of CAT's flexible competencies with the monitoring and reviewing of standards within the fixed test form process. In practice, Sireci et al. (2008) emphasized that the multistage adaptive Massachusetts Adult Proficiency Tests within adult education were introduced in 2006.

3. Item-Writing Technology

Writing techniques, whilst there have been many assessments of the style of item writing (G. H. Roid, 1986), this paper offers a summary of the existing practices related to computer-generated writing. Within the test item era, the main volume of computer submissions has focused on achievement testing and testing based on instructions.

The five key steps are pivotal in the development of an instructionally appropriate achievement test. Step one detail the definition "Conceptualization" related to the rationale behind instructions or testing (Nusche, 2013). In the first instance, a performance problem affecting students must be detected by an instructor. Similarly, a key area of content must be accepted as a direct consequence of any guideline or teaching. The analytical background of

this step is either task-related or job-related. In addition, it could be reflective. Step two emphasizes the theoretical foundation of the subject matter being taught. There are two key areas identifying the transformation of instructional commitments which can adapt into, firstly, instructional objectives that reflect the behavioral reactions by those learning, caused as a direct consequence of the teaching itself. The second area is the confirmation and detail of the subject requiring learning (G. Roid & Haladyna, 1980).

Through the use of numerous item-writing styles, a number of items can be developed. This forms the third step. The objective is to establish a platform of test items that progress the conceptual introduction from step one and transform this into an actual community of test items.

Step four is entitled item review. The key objective of this step is to recognize flawed items as well as ensure items are improved before using it in the test process. The review of items can be achieved through empirical or logical methods. These methods have been verified by Haladyna and Roid (1983).

The decision-making process which selects the items for the test is the conclusive step in the test development. These forms step five. Established modern-day test theorists, such as Hambleton et al., (1978) or Popham (1975), encourage using random item samples from the community of items, irrespective of historical guidance verifying the benefits of test blueprints and empirically-based item selection methods. Random selection is clearly evident in general theoretical content and within traditional classical conceptualizations (Brennan, 2010).

4. Classification of Item-writing Methods

The use of a ranging method to compare item-writing techniques can be anything from informal subjective to algorithmic objective. Using a topic list as an agenda for the course content or instructive writing by the teacher detailing the subject areas to be covered in a course are examples of informal methods. The following levels of the grading scale, levels two and three, offer a detailed range of ideas and goals as part of the course guidelines. These form the basis for creating a larger range of test items. According to Roid and Haladyna (1980), the scale's fourth and fifth levels illustrate a potential domain specification or an overall generic test item which can be created for the instructional course and all the related testing mechanisms.

5. Domain Specification: item progression

Introduced in 1968 by Hively et al. (1968) and Osburn (1968), "domain-referenced" testing became the subject of further development by Hively and his team (Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S., 1973). Unlike objective-based testing, which originates within instructional concepts, domain-based tests are founded within content specifications (as cited by; Haladyna & Rodriguez, 2013).

Domain specification has now been associated with a new form of technology that offers another option from objective-based item-writing techniques. Domain specification presents

five diverse approaches relating to item creation: (1) item forms, (2) linguistic-based approaches, (3) facet theory, (4) concept-based testing, and (5) computer-based methods. The definition of each approach is detailed. Thereafter, a review of any related research aligning itself to the success of method implementation, in terms of achievement testing is presented. In addition, restrictions for each approach are described (Roid, G. & Haladyna, T., 1980).

5.1. Item Forms

Specifications detailing the format and the exact terminology related to the final items form the foundation of domain-based test items (Rivera, 2006). Hively (1974) refers to specifications as "item forms." The target domain requiring assessment is one created by the community of items that ultimately create the item form. An item form presents specific characteristics. These include generation of items with a fixed syntactical structure, in excess of a single variable element, and, finally, the definition of a range of item sentences which is achieved by highlighting the replacement sets for variables". Science and Mathematics were the chosen subject areas for the creation of Hively and Osburn's item forms (as cited by; Rivera, 2006).

As noted by Bennett (2010), there are three generations of CBA's developmental level that can help evaluate 21st-century skills (Molnár, Greiff, Wustenberg, & Fischer, 2017). Compared to the PP testing, there are minor item formatting and design differences in first-generation CBA, which also uses technology to an extent (Bennett, 2015). Here, the emphasis is on item creation as well as item banking, and this generation has a positive impact on technology as it provides measures adaptively, which means the student's competency level is taken into account when selecting the next item.

The second-generation implements fewer types of traditional items and formats, such as multimedia motivators, constructed response, static performance tasks, including essays, along with developing initial tries to assess new constructs. Further, such tests also improve efficiency by using automatic item creation, as noted by Gierl and Haladyna (2012), as well as utilizing the internet regarding item review. Moreover, unique item types can also be involved as they are not similar to the traditional multiple-choice items, are visually attractive, or interesting as they focus on main competencies which are difficult to measure (Bennett, 2015).

Further, the third generation uses complicated simulations, and interactive performance tasks in which the actual environments' characteristics are replicated because of there is higher and natural interaction with computers, while evaluating unique skills in advanced manners (Bennett, 2015), including assessing dynamic problem-solving (Molnár et al., 2017).

5.2. Linguistic based approach

In order to review learning from prose, Bormuth (1970) developed the initial description identifying technology for item writing. The rules are represented in the form of a range of instructions that guide an item writer towards the best ways of revising sections of prose into a question form (as cited by; Pitoniak, 2002). According to Bormuth (1970), two forms of transformation exist (1) sentence-based items, and (2) the relations between sentences

forming the evolution of items. Illustrations of sentence-derived items can be created by the "wh-transformation" and review the recollection of the prose. The writing practice associated with these items is aligned to specific framework of regulations: "From the teaching or instructional material, it is necessary to replace any "wh" words from a sentence (who, what, where) with an appropriate alternative word, such as a noun (as cited by; Pitoniak, 2002).

5.3. Facet Theory

With an established history, structural facet theory (Foa, 1965) has primarily been a research resource, with a focus on attitude measurement. However, there has been a recent increase in applications for achievement test construction. The key benefit of this theory is the theoretical semantic references within the content analysis, as well as the absence of a need to perform empirical analysis in an attempt to retrieve a definition (Guttman & Greenbaum, 1998).

Limitations of the domain and the framework of its overall subsections are dictated by facet theory. Content structure and statistical structure are the two areas that are conceptualized and summarized. The mapping sentence is the framework that details the domain alongside the content (Guttman & Greenbaum, 1998).

5.4. Concept-Based Testing

The development of domain-based tests to address theoretical learning is based on research by Markle and Tiemann, who examined teaching and testing of theories (Markle, 1975; Tiemann, Kroeker, & Markle, 1977; Tiemann, Kroeker & Markle, Note 9). The term concept describes groups of items, events, or networks which are diverse yet have a connection under the same umbrella subcategory and with the same group name. Appreciation and comprehension of theories can be tested through a review of a generic approach towards new cases and prejudice against areas without an example.

5.5. Computer-Based Methods

Group storage of item banks was the basis for original efforts in computer-based testing. Samples from these item banks were then used for testing (Cantillon, Irish, & Sales, 2004). Alternative systems adopt true item formation, an example being the Stanford Computer-Assisted Instruction (CAI) project's mathematics and reading applications (Roid & Haladyna, 1980). The creation of an item through a computer program can potentially develop domain-based tests. These are illustrated through flow chart diagrams or program listings.

Roid promoted the use of a number of CAI languages that contained specific functions which enable the generation of items. Such languages include COURSEWRITER, PLANIT, and TUTOR. COURSEWRITER was used to develop criterion examinations and for creating reading projects. For example, a pilot exercise included phrases such as "Bill wore the hat." The student sitting the exam would be required to complete the sentence by finding missing words from a range of words assembled at random by the computer. Such words would include "tan," "fat," "man," or "run." The responses are obtained from the guidance received from words used in the prior reading program (Roid, 1986).

Based on research reviewing misspellings by writers of spelling items, Fremer and Anastasio (1969), proactively used computers to create spelling test items (as cited by; Roid, 1986).

6. The Advantages and Disadvantages of Computer-Based Systems

Computer-based assessment is valuable in education because young people are deemed to be (Wantulok, 2015) . In other words, they possess a far greater familiarity with technology than do adults, and thus classroom engagement with technology can be used to facilitate learning and the acquisition of multi-tasking proficiencies (Wantulok, 2015). In addition, it considers a motivational factor to learn.

Technology can be employed in various ways to develop an educational assessment. It can help to record students' cognitive, psychomotor, and affective features with greater efficiency as well as the social contexts involved in the learning and teaching processes. Moreover, technology provides a more accurate and measurable method to gather information and process it while allowing enhanced data analysis. Such enhanced data analysis ensured that feedback is provided accurately and quickly to participants as well as stakeholders, thus effectively validating the decision-making process (Casapò, Ainley, Bennett, Latour, & Law, 2012).

CBT shows to have considerable advantages over paper and pencil test, for both who are run the assessment process and for the students who are test-takers. These advantages are identified by the U.S. Department of Education. There currently is increasing interest in Computer-Based Test and advocates have recognized numerous positive merits of this method of assessment including, efficient administration and management, individual preference, self-selection options for students, enhancing writing performance, quickly getting results, efficient item development, risen authenticity, and the potential to shift attention or focus from assessment to instruction (Thurlow, Lazarus, Albus, & Hodgson, 2010).

Furthermore, technology-assisted assessment can reduce the amount of time, resources, or disruption involved in administering evaluations. Such assessments can additionally create a finer impression regarding students' abilities, requirements, and interests than is possible through traditional testing. Thus, individualized learning possibilities become a realizable objective (Gohl, Gohl, & Wolf, 2009).

Using technology for assessments permits a multitude of enhanced formats far exceeding basic true or false, multiple-choice, or gap-fill type questions (Office of Educational Technology). One example of an enhanced technique is the graphic response, which requires a response comprising drawing, moving, or choosing graphic elements. Conversely, equation response necessitates submitting an equation. Other examples include hot text, in which students reposition sentences within the text, and performance-based assessment wherein students completed several complex tasks (Culatta, 2016).

In essence, students can use technology-based assessment to illustrate more complex cognitive processes and reveal more sophisticated comprehension in regard to the material than is permitted by conventional testing.

Such assessment requires that students demonstrate a complex skillset, fusing and analyzing data from manifold sources, and rationalizing any deductions. Thus, an English language assessment task might involve analyzing original texts and writing an essay in response. A mathematics performance task could require the student to outline the linear relationship between quantities on a graph which has been based on real data. Performance-based evaluations depend on students creating their own response, not simply selected from pre-determined optional responses. Knowledge must be applied to explain authentic problems (Darling-Hammond, Adamson, & Abedi, 2010).

In some cases, technology can be utilized to permit students engaged in performance-based assessment to enter responses online and can be combined where necessary with hand-scoring to present a total test score. This can be seen with the Partnership for Assessment of Readiness for College and Careers and the Smarter Balanced Assessment Consortium evaluation, where machine scoring is combined with classroom speaking and listening assessment (Culatta, 2016).

CBT surpasses traditional paper-based testing in that each student can take a bespoke test, unique to their abilities. The computer reacts to each examinee's performance, avoiding questions that are too simple or too difficult, eschewing scenarios which fail to add to the assessment. All examinees will find their test equally stimulating yet not beyond their scope (Yunxiang, Ruixue, Lili, Qiao, & Hefei, 2010).

Hence, a correctly answered question leads to a more difficult follow-up question, whereas an incorrectly answered question is met with an easier one. Overall scores will be based on an amalgamation of correct answers and question level. Thus, questions carry different weightings that impact the final scores (Yunxiang, Ruixue, Lili, Qiao, & Hefei, 2010).

According to Mazzeo and Harvey (1988), computer-based scoring is just as susceptible to distortion as paper-based testing. Moreover, they add, reading passages presented on computer screens can create problems too. Bunderson, Inouye, and Olsen (1988) similarly highlight concerns over slower paragraph comprehension, which results from computer presentation. However, tasks such as coding have been shown to take less time on screens. Furthermore, Kuzmina (2010) suggests that computers possess a far greater capacity to present assessment materials than do paper-based tests. This is exemplified in motion effects, rotating geometric shapes, animated trajectories, 3-D diagrams, and plants presented from multiple angles.

Computer-based testing can both facilitate individualized testing and permit it to be conducted at stages most suitable for each student. Testing can be bespoke, tailored for each student, and flexible therein (Chou, 2000). Moreover, it can provide rapid feedback and scoring. Yet, according to Kimlzuka and Taniguchi (1986), computer-based testing results may be skewed due to associated anxiety, which they regard as an essential issue requiring attention. Specifically, Wise and Plake (1989) identify the instant feedback provided by computer-based assessment as a major anxiety-inducing issue for students. Hence, Jacob and Chase's suggestion that item by item feedback is halted until further research has clarified its impact (Kuzmina, 2010).

This paper will include several comparisons provided in previous studies between CBT and PBT for supporting the aforementioned statements. The first article is Jeong's 2014 paper titled "A comparative study of scores on computer-based tests and paper-based tests" that compared the numerous Korean students' CBT and PBT versions of the same test. Surprisingly, although the Korean students who participated in the study were more familiar to use technologies like computers, the Internet, and multimedia than students, they did not achieve higher results than the CBT test results from PBT. The study's results showed that the participants scored higher in PBT in all subjects compared to CBT and that there were significant differences in the two subject's results: Korean language ($F= 25.612$, $p < 0.01$); science ($F= 6.386$, $p < 0.05$). On the other hand, the CBT as well as PBT test scores in mathematics ($F = 2.077$, $p > 0.05$) did not differ significantly from that in social studies ($F = 1.111$, $p > 0.05$).

The article also highlights the differences in test grades of male and female participants for different subjects. In terms of the average CBT and PBT results, it was noted that male students' results differed for every subject, with Korean language subject showing the most difference ($F = 13.980$, $p < 0.01$), while the difference for the CBT and PBT results were not significant in mathematics ($F = 0.084$, $p > 0.05$), science ($F = 1.866$, $p > 0.05$), and social studies ($F = 0.237$, $p > 0.05$).

Compared to the male students' results, it was noted that the female students' scores varied significantly in three subjects which included Korean language ($F = 11.370$, $p < 0.01$), science ($F = 5.199$, $p < 0.05$), and mathematics ($F = 4.042$, $p < 0.05$). However, their social studies test scores in CBT as well as PBT did not differ significantly ($F = 1.014$, $p > 0.05$).

The other study is Yunxiang, Ruixue, Lili, Qiao, and Hefei's "Advantages and disadvantages of computer-based testing: a case study" published in 2010. These authors presented CBT's benefits as well as drawbacks by implementing the case study for examining CBT's scoring algorithm. This study involved experimental graduates regarding service-learning who were provided with certain CBT pre-tests and post-tests. As seen in the quantitative results, the subjects' English language performance improved, which was statistically important, considering the quantitative signs. Further, the service-learning project participants' scores were 4 points higher than in their pre-test.

As per this result, the students involved in the service-learning perform showed improved learning. This indicates that learning achievement has a connection to the learners' experience of real-life language usage and is based on social communication. Therefore, this validates Ellis' belief in social communication or natural settings resulting in improved L2 proficiency compared to formal institutional environments (Ellis, 1994).

Karadeniz (2009) also evaluated how students' achievements are influenced by paper-based, mobile-based, and web-based assessments. This study was conducted for three weeks and involved 38 students. It was observed that there were significant differences between students' scores in the study's second week compared to the first week. Moreover, students began having a more positive attitude concerning web-based and mobile-based assessment as it was comprehensive, easier to use and adapt to, and providing quick feedback. In addition,

web-based tests were more favored, while paper-based ones were least favored.

Bodmann and Robinson (2004) also conducted a study in which they compared the performance and speed between CBTs and PPTs. There were 30 MCQs items in CBTs as well as PPTs, and both tests were to be completed in 35 minutes. It was observed that 28 students, which is almost half the class, first took the test on the computer, while the other students took the paper version first. After two weeks, the first and second group of students received PPTs and CBTs, respectively. It was noted that undergraduates' performance was quicker in CBT compared to PBT, while the results remained the same (Ghaderi, Mogholi, & Soori, 2014).

Benefits and drawbacks of PBT and CBT:

6.1. Costs

As noted by Farcot and Latour (2009), a substantial amount is required to purchase equipment, including computers, infrastructure, and laboratory tools for conducting technology-based examinations. This problem is experienced by colleges and institutes. It should be noted that implementing CBTs reduces mid-term as well as long-term costs (Ben\Ho Csapó & Molnár, 2017; Farcot & Latour, 2009; Kuzmina, 2010).

6.2. Speed and safety of the data flow

CBTs ensure faster and easier data processing (Csapó & Molnár, 2017; Csapó, Lörincz, & Molnár, 2012) and is a safer method of ensuring security as it enables the use of username and password (Kuzmina, 2010; Marriott & Teoh, 2012). Further, as one can choose questions randomly, the possibility of cheating is decreased while objectivity is enhanced (Marriott & Teoh, 2012).

6.3. Developments in the reliability of the tests

Systems such as computer adaptive testing provide students with assignments and questions as per the test standard while also providing the response time provided for the ideal items. This will lead to more reliable PBTs (Molnár & Csapó, 2018). Meanwhile, CBA can ensure validity by assessing and scoring tests as well as item-level data automatically, along with including in-depth feedback concerning students' tests, task/question-level performances, and subtests (Marriott & Teoh, 2012).

6.4. Possibilities in adaptive testing

When new technology is adopted in adaptive testing, the real item is presented that is according to the success of the applicant when solving the earlier item. Thus, considering CBT, as stated by Benö Csapó et al. (2012), the applicant's score is taken into account when determining the time and decision concerning the next step. In this approach, the students are established when starting the test.

6.5. Changes in students' motivation for testing

It is important to note that technology provides new opportunities in terms of education. Technology encourages learning as it enables a creative task to be presented using multimedia, thus increasing enjoyment as well as motivation (Molnár & Horincz, 2012). When conducting an assessment, it ensures that the environments entertaining as students tend to gravitate towards them (Molnár & Horincz, 2012; Ridgway, McCusker, & Pead, 2004). Moreover, Kuzmina (2010) noted that CBA tends to be used more by a specific student group such as those with a disability, as PP tests could be limiting for them (Kuzmina, 2010). Hence, using CBA may motivate them for testing.

6.6. Possibilities for assessing new constructs

As noted by Alruwais, Wills, and Wald (2018), CBA has paved the way to develop and implement more complex, unique, innovative items apart from the traditional first-generation computer-based items such as multiple-choice. Compared to the first-generation items, the second-generation items ensured that there were multimedia elements that developed more real-life problems as well as a better-standardized testing environment, such as where each individual will be able to listen to the same voice. Further, the third-generation tests (Ripley, Harding, Redif, & Britain, 2009; Greiff, Wüstenberg, & Funke, 2012) such as simulations, cooperation, and interaction (see Csapó & Molnár, 2017) enabled assessing construct that may have been impossible to measure using traditional assessments and standard item formats (e.g., Complex Problem Solving (CPS); see Greiff et al., 2012; Dörner & Funke, 2017 - in PISA 2012, known as Creative Problem Solving). Thus, using second-generation and third-generation tests can lead to replicating complex real-life situations, utilizing authentic tasks, dynamism, interactions, inter-test collaboration, and virtual worlds (Molnár & Csapó, 2019; Molnár et al., 2017) for assessing more complex 21st-century skills (Pachler, Cook, & Bachmair, 2010; Molnár et al., 2017; Molnár & Csapó, 2018; Ridgway et al., 2004).

7. Conclusion

The information and communication technology (ICT) development has not only revolutionized society and provided unique tools, as noted by Molnár and Csapó (2019), but also offered new possibilities as well as challenges concerning educational assessment. As noted by Molnar and Csapó (2018), new assessments are required for measuring as well as developing skills of the 21st century. Such assessment methods must surpass testing the actual knowledge while provides prompt and meaningful feedback for teachers and learners. Traditional assessment methods cannot help with this issue (Molnár & Csapó, 2019).

There are three steps involved in this development that can lead to the continuously expanding possibilities concerning educational assessment. While the first-generation CBA tests appear to be similar to the traditional PP test, CBA offers numerous advantages such as feedback time and delivery mode. Moreover, the second-generation CBA involves multimedia aspects that enable adaptive testing. The third-generation tasks can help with measuring more complicated constructs, including 21st-century skills through simulation, interaction, dynamically changing items, and cooperation (see Csapó & Molnár, 2017).

CBA was being regarded as a feasible substitute for PP testing during the 20th century concerning large-scale assessments. Although the numerous media studies that have been conducted have shown diverse results because of different samples, skills and abilities, assessed knowledge, and item formats being used, extensive studies have been conducted focusing on PP's differences from the CB delivery mode as well as students' test performance. Recent studies have shown that PP testing can be compared with CB testing. Although there are differences, they can diminish with time when computers start being available widely, and students would be more attracted to CB tests rather than PP testing. Hence, the impacts of test mode are no longer a problem, and thus, the focus can be shifted to other possibilities of the new technologies in terms of educational assessment.

Using technology has significantly enhanced the testing procedures' efficiency by making data collection faster, accelerating data processing, supporting real-time automatic scoring, providing immediate feedback, and transforming the entire assessment process including innovative task presentation (see Csapó, Lőrincz, & Molnár, 2012 for an in-depth examination of technological issues). Further, it led to unique possibilities concerning the item and test development. Technology also helps store and analyze contextual data in new ways, called educational data mining, learning analytics, or logfile analyses, which is slightly different from the analyses. These multiple benefits indicate that significant assessments in the future will be conducted using a technological environment. This is evident in the major international large-scale summative assessments (e.g., IEA TIMSS, PIRLS; OECD PISA). In the last few years, using one of the larges possibilities of CBA, the automatic feedback, there has been an emphasis on individualized diagnostic assessment beyond the mainly summative approach, thus using the power of proper and prompt feedback information to personalize learning and instruction (Molnár & Csapó, 2019).

The fact that CBA has replaced PP at every testing level, including formative or summative and high or low stakes, while providing new opportunities in the assessment field in the form of online diagnostic assessment, embedded assessment, adaptive testing, gaining more information about the students' test-taking behavior through assessing log files, and measuring new constructs cannot be denied. As stated by Csapó et al. (2012), this technology also broadens the possibilities from qualitative as well as quantitative perspectives while strengthening CBA's use.

References

Alruwais, N., Wills, G., & Wald, M. (2018). Advantages and challenges of using e-assessment. *International Journal of Information and Education Technology, 8*(1), 34-37.

Bennett, R. (2010). Innovative assessment systems: The role of new technology. https://doi.org/10.18178/ijiet.2018.8.1.1008

Bennett, R. E. (2015). The changing nature of educational assessment. R*eview of Research in Education, 39*(1), 370-407. https://doi.org/10.3102/0091732X14554179

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

51-60. https://doi.org/10.2190/GRQQ-YT0F-7LKB-F033

Bormuth, J. R. (1970). On the Theory of Achievement Test Items: With an Appendix" On the Linguistic Bases of the Theory of Writing Items", by P. Menzel. University of Chicago Press.

Breithaupt, K., Ariel, A. A., & Hare, D. R. (2009). Assembling an inventory of multistage adaptive testing systems. In *Elements of adaptive testing* (pp. 247-266). Springer. https://doi.org/10.1007/978-0-387-85461-8_13

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1-21. https://doi.org/10.1080/08957347.2011.532417

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). The four generations of computerized educational measurement. *ETS Research Report Series, 1988*(1), i--148. https://doi.org/10.1002/j.2330-8516.1988.tb00291.x

Cantillon, P., Irish, B., & Sales, D. (2004). Using computers for assessment in medicine. B*mj, 329*(7466), 606-609. https://doi.org/10.1136/bmj.329.7466.606

Casapò, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological Issues for Computer-Based Assessment. I P. Griffin, B. McGaw, & E. Care (Red.). *Assessment and Teaching of 21 St Century Skills,* 143-230. https://doi.org/10.1007/978-94-007-2324-5_4

Chou, C. (2000). Constructing a computer-assisted testing and evaluation system on the World Wide Web-the CATES experience. *IEEE Transactions on Education, 43*(3), 266-272. https://doi.org/10.1109/13.865199

Clyman, S. G., Melnick, D. E., & Clauser, B. E. (1995). Computer-based case simulations. *Assessing Clinical Reasoning: The Oral Examination and Alternative Methods,* 139-149.

College Board, A. (2017). Administrator ’ s Manual.

Csapó, Ben\Ho, & Molnár, G. (2017). Potential for assessing dynamic problem-solving at the beginning of higher education studies. *Frontiers in Psychology, 8,* 2022. https://doi.org/10.3389/fpsyg.2017.02022

Csapó, Benö, Lörincz, A., & Molnár, G. (2012). Innovative assessment technologies in educational games designed for young students. In *Assessment in game-based learning* (pp. 235-254). Springer. https://doi.org/10.1007/978-1-4614-3546-4_13

Darling-Hammond, L., Adamson, F., & Abedi, J. (2010). Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning. Stanford Center for Opportunity Policy in Education.

DeVore, R. (2002). Considerations in the development of accounting simulations. Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Dikusar, A. (August 9, 2018). The Use Of Technology In Special Education - eLearning Industry. Retrieved January 30, 2020, from https://elearningindustry.com/use-of-technology-in-special-education

Dörner, D., & Funke, J. (2017). Complex problem solving: what it is and what it is not. *Frontiers in Psychology, 8,* 1153. https://doi.org/10.3389/fpsyg.2017.01153

Ejim, S. (2018). An Overview of Computer Based Test. (April). https://doi.org/10.13140/RG.2.2.32040.88326

Ellis, R. (1994). The study of second language learning. *Social Factors and Second Language Learning, 6,* 197-210.

Farcot, M., & Latour, T. (2009). Transitioning to computer-based assessments: A question of costs. *The Transition to Computer-Based Assessment,* 108-116.

Foa, U. G. (1965). New developments in facet design and analysis. *Psychological Review, 72*(4), 262. https://doi.org/10.1037/h0021968

Fremer, J., & Anastasio, E. J. (1969). COMPUTER-ASSISTED ITEM WRITING—1 (SPELLING ITEMS) 1. *Journal of Educational Measurement, 6*(2), 69-74. https://doi.org/10.1111/j.1745-3984.1969.tb00661.x

Ghaderi, M., Mogholi, M., & Soori, A. (2014). Comparing between computer based tests and paper-and-pencil based tests. *International Journal of Education and Literacy Studies, 2*(4), 36-38. https://doi.org/10.7575/aiac.ijels.v.2n.4p.36

Gierl, M. J., & Haladyna, T. M. (2012). A*utomatic item generation: Theory and practice.* Routledge. https://doi.org/10.4324/9780203803912

Gohl, E. M., Gohl, D., & Wolf, M. A. (2009). Assessments and technology: A powerful combination for improving teaching and learning. Meaningful Measurement: The Role of Assessments in Improving High School Education in the Twenty-First Century. Washington, DC: Alliance for Excellent Education.

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement, 36*(3), 189-213. https://doi.org/10.1177/0146621612439620

Guttman, R., & Greenbaum, C. W. (1998). Facet theory: Its development and current status. *European Psychologist, 3*(1), 13-36. https://doi.org/10.1027/1016-9040.3.1.13

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items.* Routledge. https://doi.org/10.4324/9780203850381

Haladyna, T. M., & Roid, G. H. (1983). Reviewing criterion-referenced test items. *Educational Technology, 23*(8), 35-38.

Hively, W. (1974). Introduction to domain-referenced testing. *Educational Technology, 14*(6), 5-10.

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology, 33*(4), 410-422. https://doi.org/10.1080/0144929X.2012.710647

Karadeniz, S. (2009). The impacts of paper, web and mobile based assessment on students' achievement and perceptions. Scientific Research and Essay, 4(10), 984-991.

Kimlzuka, N., & Taniguchi, M. (1986). *No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析. Title. 34*(10), 1986.

Kraslawski, A., & Turunen, I. (2013). *23rd European Symposium on Computer Aided Process Engineering.* Elsevier.

Kuzmina, I. P. (2010). Computer-based testing: advantages and disadvantages. Вісник Національного Технічного Університету України Київський Політехнічний Інститут. Філософія. *Психологія. Педагогіка,* (1), 192-196.

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education, 19*(3), 189-202. https://doi.org/10.1207/s15324818ame1903_2

Luecht, R. M., & Sireci, S. G. (2011). A Review of Models for Computer-Based Testing. *Research Report 2011-12.* College Board.

Markle, S. M. (1975). They teach concepts, don't they? *Educational Researcher, 4*(6), 3-9. https://doi.org/10.3102/0013189X004006003

Marriott, P., & Teoh, L. (2012). ICT for assessment and feedback on undergraduate accounting modules. The Higher Education Academy. Available from Http://Www. Heacademy. Ac. Uk/Resources/Detail/Disciplines/Finance-and-Accounting/Using-ICT-in-Assessment-and-Feedback.

Mazzeo, J., & Harvey, A. L. (1988). The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature. *ETS Research Report Series, 1*, i--27. https://doi.org/10.1002/j.2330-8516.1988.tb00277.x

McCain, G. (1995). Technology-based assessment in special education. *THE Journal (Technological Horizons In Education), 23*(1), 57.

Melican, G. J., Breithaupt, K., & Zhang, Y. (2009). *Designing and implementing a multistage adaptive test: the uniform CPA exam. In Elements of adaptive testing* (pp. 167-189). Springer. https://doi.org/10.1007/978-0-387-85461-8_9

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*(4), 287-304. https://doi.org/10.1207/s15324818ame0904_1

Molnár, G., & Csapó, B. (2018). The efficacy and development of students' problem-solving strategies during compulsory schooling: Logfile analyses. *Frontiers in Psychology, 9,* 302. https://doi.org/10.3389/fpsyg.2018.00302

Molnár, G., & Csapó, B. (2019). Technology-based diagnostic assessments for identifying

early mathematical learning difficulties. In *International handbook of mathematical learning difficulties* (pp. 683-707). Springer. https://doi.org/10.1007/978-3-319-97148-3_40

Molnár, G., Greiff, S., Wustenberg, S., & Fischer, A. (2017). *Empirical study of computer based assessment of complex problem solving skills.* https://doi.org/10.1787/9789264273955-10-en

Molnár, G., & L\Horincz, A. (2012). Innovative assessment technologies: Comparing 'face-to-face'and game-based development of thinking skills in classroom settings. International Proceedings of Economics Development and Research. *Management and Education Innovation, 37*, 150-154.

Nusche, R. (2013). Student assessment: Putting the learner at the centre. Synergies for Better Learning: An International Perspective on Evaluation. Chapter 4, 139-270. Reviews of Evaluation and Assessment in Education and Assessment. OECD Publishing, Paris. https://doi.org/10.1787/9789264190658-7-en

Pachler, N., Cook, J., & Bachmair, B. (2010). Appropriation of mobile cultural resources for learning. *International Journal of Mobile and Blended Learning (IJMBL), 2*(1), 1-21. https://doi.org/10.4018/jmbl.2010010101

Pitoniak, M. J. (2002). Automatic Item Generation Methodology in Theory and Practice1, 2.

Alliance, R. (Oct 23, 2017). Importance of Technology in assessing students learning profiles. Retrieved January 30, 2020, from https://medium.com/@readalliance/edtech-a-shift-from-paper-based-to-a-more-tech-enabled-process-of-assessing-students-2d30093c134d

Culatta, R. (January 2016). Assessment - Office of Educational Technology. Retrieved January 30, 2020, from https://tech.ed.gov/netp/assessment/

Ridgway, J., McCusker, S., & Pead, D. (2004). *Literature review of e-assessment.*

Ripley, M., Harding, R., Redif, H., & Britain)(JISC), J. I. S. C. (Great. (2009). *Review of Advanced e-Assessment Techniques* (RAeAT).

Rivera, J. (2006). Test item construction and validation: Developing a statewide assessment for agricultural science education.

Roid, G. H. (1986). 3. Computer Technology In Testing.

Roid, G., & Haladyna, T. (1980). The emergence of an item-writing technology. *Review of Educational Research, 50*(2), 293-314. https://doi.org/10.3102/00346543050002293

Roschelle, J. M., Pea, R. D., Hoadley, C. M., Gordin, D. N., & Means, B. M. (2000). Changing how and what children learn in school with computer-based technologies. *The Future of Children*, 76-101. https://doi.org/10.2307/1602690

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation*. American Psychological Association. https://doi.org/10.1037/10244-000

Sireci, S. G., Baldwin, P., Martone, A., Zenisky, A. L., Kaira, L., Lam, W., … others. (2008). *Massachusetts Adult Proficiency Tests Technical Manual,* Version 2. Center for Educational Assessment Research Report No, 677.

Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). Computer-Based Testing: Practices and Considerations. Synthesis Report 78. National Center on Educational Outcomes, University of Minnesota.

Tiemann, P. W., Kroeker, L. D., & Markle, S. M. (1977). Teaching verbally mediated coordinate concepts in an ongoing college course. Annual Meeting of the American Educational Research Association, New York.

Wantulok, T. (February 12, 2015). How Important is Technology in Education? Pine Cove's Top 10 Reasons. Retrieved January 30, 2020, from https://marketing.pinecc.com/blog/the-importance-of-technology-in-education-pine-coves-top-10-reasons

Wise, S. L., & Plake, B. S. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice, 8*(3), 5-10. https://doi.org/10.1111/j.1745-3992.1989.tb00324.x

Yunxiang, L., Ruixue, G., Lili, R., Qiao, W., and Hefei, Q. (2010). Advantages and disadvantages of Computer-based Testing: A case study of service learning. *The 2nd International Conference on Information Science and Engineering,* 1-4. https://doi.org/10.1109/ICISE.2010.5691870

Copyright Disclaimer