The evolution of technology-based assessment: past, present, and future

Saleh Ahmad Alrababah*

Doctoral School of Education,
University of Szeged, Institute of Education,
H-6722 Szeged, Petőfi Sándor sgt. 30-34, Hungary
Email: alrababah.saleh.ahmad@edu.u-szeged.hu
*Corresponding author

Gyöngyvér Molnár

Department of Learning and Instruction,
MTA–SZTE Research Group on the Development of Competencies,
Center for Research on Learning and Instruction,
University of Szeged, Institute of Education,
H-6722 Szeged, Petőfi Sándor sgt. 30-34, Hungary
Email: gymolnar@edpsy.u-szeged.hu

Abstract: This paper presents developmental trends in technology-based assessment in an educational context and highlights how technology-based assessment has reshaped the purpose of educational assessment and the way we think about it. Developments in technology-based assessment stretch back three decades. Around the turn of the millennium, studies centred on computer-based and paper-and-pencil test comparability to ascertain the effect of delivery medium on students' test achievement. A systematic review of media studies was conducted to detect these effects; the results were varied. Recent work has focused on logfile analysis, educational data mining and learning analytics. Developments in IT have made it possible to design different assessments, thus boosting the number of ways students can demonstrate their skills and abilities. Parallel to these advances, the focus of technology-based assessment has shifted from an individual and summative approach to one which is cooperative, diagnostic and more learning-centred to implement efficient testing for personalised learning.

Keywords: information and communications technology; ICT; computer-based assessment; CBA; personalisation of instruction; time on task; media comparison studies.

Reference to this paper should be made as follows: Alrababah, S.A. and Molnár, G. (2021) 'The evolution of technology-based assessment: past, present, and future', *Int. J. Learning Technology*, Vol. 16, No. 2, pp.134–157.

Biographical notes: Saleh Ahmad Alrababah is a PhD candidate at University of Szeged, Institute of Education. His areas of interests are about technology-based assessment and its inventions, and ICT in education.

Gyöngyvér Molnár is a Professor of Education and the head of Institute of Education. She chairs the Hungarian Academy of Sciences Education and Psychology Committee in Szeged. Her main areas of interest include:

technology-based assessment, improving cognitive skills, and the potential for using ICT in education – all of which are aimed at improving the quality of learning. She heads eDia, an online diagnostic assessment system used in numerous countries. She serves on the editorial boards of numerous domestic and international journals. She received the Apáczai Csere János Prize, a ministerial acknowledgement for her outstanding scholarly work in support of educational practice in 2016.

1 Introduction

Paper-based (PB) testing, which falls under 'traditional assessment', has played a key role in educational assessment. Its possibilities are greatly restricted compared to technology-based assessment (TBA). TBA covers all forms of assessment which are delivered and marked with the aid of technology, that is, via the most commonly used computers [computer-based assessment (CBA)] or other electronic tools and devices (Kuzmina, 2010). In other words, through TBA there is an interaction between the student and the technology used. We are aware that computers play a dominant role in TBA because of its versatility. We have thus decided to use these terms as synonyms in the study.

If CBA is delivered online, which is the main focus of the present discussion, the benefits increase significantly, mostly building on the possibilities of automatic scoring and feedback. Other forms of TBA and CBA (e.g., optical mark readers for multiple-choice tests) are excluded from the main discussion.

Traditional paper-and-pencil (PP) tests are usually fixed tests; thus, every student receives the same items and tasks in the same order during data collection, independent of ability level. The most crucial disadvantages of PP tests are the long feedback time, the restricted suitability of test design, including difficulty, and the use of a limited range of item types.

The use of technology in assessment may lead to improved assessment, thus offering numerous advantages (e.g., automatic item generation, presenting dynamic stimuli and automatic scoring; Becker, 2004; Csapó et al., 2014; Dikli, 2006; Mitchell et al., 2002; Valenti et al., 2003), cutting costs (e.g., delivery, distributing results and evaluating answers; Bennett, 2003; Christakoudiset al., 2011; Wise and Plake, 1990) and laying the groundwork for new innovations (e.g., measuring new constructs and using new item types; Dörner and Funke, 2017; Pachler et al., 2010) in educational assessment. The possibilities, advantages and challenges of TBA are growing in accordance with the level of application (e.g., item development, delivery, scoring and feedback), type of technology (e.g., desktop computer, touchscreen tablets and eye-tracking technologies), methodology used (e.g., fixed testing or adaptive testing), delivery (e.g., internet-based, local server delivery and delivery on removable media), scoring (e.g., automatic, computer-based (CB), but not automatic, human scoring; item-level scoring based on the actual answer of the students or logfile and process data analyses based on the actions of the students), item types (e.g., traditional multiple-choice or state-of-the-art third-generation innovative item types, including interactivity), domains (e.g., domains can be assessed using traditional methods, such as reading fixed texts, or domains requiring TBA, such as reading digital and printed texts) and the technological conditions

of the assessment. Through technology, teachers and educational authorities and managers can develop new policies that truly meet the expectations of the 21st century (Shatunova et al., 2019), e.g., measuring 21st century skills [i.e., critical thinking, problem-solving, creativity, collaboration (teamwork), learning to learn, entrepreneurship and information literacy (Binkley et al., 2012; Redecker et al., 2010)] even on international large-scale assessments (LSA) [see e.g., the OECD Programme for International Student Assessment (PISA) creative or collaborative problem-solving module; Griffin et al., 2012; OECD, 2014).

Information and communications technologies, especially computers, have had an immense impact on the development of educational assessment not only from a quantitative perspective, but also from a qualitative one. New science has emerged in educational assessment, which focuses not only on an analysis of the actual answer and achievement data, but more deeply on an analysis of contextual data gathered during data collection beyond the actual answers provided by the students. Logfile analysis, educational data mining and learning analytics (Csapó et al., 2014; Johnson et al., 2016; Wise, 2019) have become the state of the art in educational assessment analysis and attracted increasing research interest. They make it possible to answer research questions that would be unanswerable using traditional assessment techniques.

To sum up, this paper presents a systematic literature review of the different qualitative or quantitative stages in the development of TBA, from the first use to the latest developments, including a systematic analysis of the media effect and media comparison studies on students' performance using the same test (or measuring the same construct) in different media. We also present and discuss the impact of large-scale international assessments on the evolution of TBA and the challenges of TBA developments for the future.

2 Research questions

We posited the following research questions on developmental trends in CBA:

RQ1 What role does technology play in educational assessment?

RQ2 Do large-scale international assessments have on effect on the evolution of TBA? If so, what is the nature of this effect?

RQ3 Are PP and CB test results comparable?

RQ4 What is required for the application of CBA among kindergarten children and its systematic integration into everyday school practice?

RQ5 How can an advanced use of the advantages and possibilities of TBA promote a shift in the aim of assessment from effective summative testing to personalised learning?

3 Early studies in TBA

Using technology in assessment started in the 1920s when Sidney L. Presses designed a machine for testing (Alruwais et al., 2018; Skinner, 1958). 1935 saw the first attempt to

use a test scoring machine, the IBM model 805, to test millions of Americans in a type of objective test (Khoshsima and Hashemi, 2017). In the 1970s and 1980s, new computer systems were launched in language testing for purposes (test design, test construction, tryout, delivery, management, scoring, analysis and interpretation, and reporting) beyond simple test scoring (Fulcher, 2000).

The next major development took place in the 1990s, with the focus on the applicability of a broad range of technologies from the most common to the cutting edge (Baker and Mayer, 1999). In recent decades, educational assessment has represented one of the most dynamically developing areas in education; as a result, CBA has become part of large-scale international assessments.

In the early studies of this implementation process, the focus was on the comparability of traditional (PP or face-to-face) and CB test results, or media comparison studies. In media comparison studies, researchers compare the test results of students tested with one medium versus those of – in an ideal case, the same – students tested with another medium using the same test or at least measuring the same construct. It is challenging to conduct valid media comparison research because of difficulties in ensuring that the results are only influenced by the test medium.

Most types of traditional items, such as multiple-choice items, could easily be transferred to a CB assessment platform. The common research question among these studies was the following: whether traditionally administered test results are equivalent to those of CB tests using the same questions and item formats for determining score equivalence (Kuzmina, 2010).

The *Guidelines for Computer-Based Tests and Interpretations* published by the American Psychological Association (APA) in 1986 specified score equivalence between CB and PP tests. They concluded that

1 the rank order of the test scores in PP and in CB mode was approximately the same

2 the means, standard deviations and shapes of the distribution curves were also nearly the same, at least after rescaling and transforming the data (APA, 1986; Kuzmina, 2010).

In parallel with this issue and building on the results of the different media studies, a great deal of research highlighted the significance and benefits of TBA over traditional PB testing.

4 The effect of large-scale national and international assessments on the evolution of TBA

Around the turn of the millennium, large-scale international assessments [e.g., the National Assessment of Educational Progress (NAEP) and Programme for International Student Assessment of the OECD (PISA)] were conducted to capitalise on CB delivery and implement TBA (Csapó et al., 2012; OECD, 2010) with the aim of replacing traditional face-to-face and PP testing. One of the hot topics of this period was a comparison of the results of PP and CB assessments for the same construct (Kingston, 2008; Wang et al., 2008).

Csapó and Molnár (2019) summarised the role of large-scale international assessment in the development of TBA. They argued that the OECD PISA assessments have had an

impact on the development of TBA in two major ways: they have advanced the technological infrastructure, and they have tested the preparedness of different countries for the assessments. In PISA the first CBA took place in 2006, when the computer-based assessment of science was an optional domain (OECD, 2010). Only three countries took part in the data collection (Denmark, Iceland and Korea), but this research served as good practice for future assessments. Three years later, an assessment of digital reading was an extra optional domain in PISA. The research design made it possible to compare the results in PP and digital reading (OECD, 2011). In the following PISA cycle, assessments for reading and mathematics as well as creative problem-solving as an innovative domain were offered in CB delivery mode (OECD, 2013, 2014). This assessment has had a huge impact on the development of CBA and has resulted in a complete shift from PP to CB testing in PISA (OECD, 2016); thus, in 2015, the transition of PISA to CBA was complete, with all the assessments being administered via computer.

The Trends in International Mathematics and Science Study (TIMSS) of the International Association for the Evaluation of Educational Achievement (IEA) is an international comparative study measuring fourth and eighth graders' achievement in mathematics and science as a continuation of IEA's previous studies conducted from the 1960s through the 1980s. Since 1995, with a four-year assessment cycle, TIMSS has assessed student achievement using PP methods on six occasions – in 1999 (eighth grade only), 2003, 2007, 2011 and 2015 (Mullis and Martin, 2017). In the 2019 assessment cycle, TIMSS shifted to CBA and was called eTIMSS with expanded problem-solving and inquiry tasks and novel item types, including drag and drop, sorting and drop-down menu input types. Just around half of the 65 TIMSS countries used eTIMSS in 2019, while the remainder administered TIMSS with the PP format. The shift from traditional PP administration to a fully CBA expanded the coverage of the TIMSS assessment frameworks (Fishbein et al., 2018).

The International Reading Literacy Study (PIRLS) is an assessment of reading comprehension in the fourth grade, which was developed by the IEA and has been conducted every five years since 2001. PIRLS provides information on trends in reading literacy achievement among students in countries that have participated in the assessment cycles. PIRLS was expanded in 2016 to include ePIRLS – an innovative assessment of online reading. ePIRLS is a CBA that uses an engaging, simulated Internet environment to present students with authentic school-like assignments involving social studies and science topics (Mullis et al., 2017).

The IEA has long been concerned with the use of information and communications technology (ICT) in education. The first IEA study in this field was the Computers in Education Study (COMPED) conducted in 1989 and 1992, followed by IEA's Second Information Technology in Education Study (SITES) Module 1 in 1998–1999 and Module 2 in 2001 and 2006, which assessed ICT goals and practices in education and the infrastructure in twenty-six countries (Fraillon et al., 2019). In 2013, the first cycle of the International Computer and Information Literacy Study (ICILS) was conducted, collecting data in 21 education systems. It investigated how students in Grade 8 in these countries developed the ICT literacy skills that would enable them to participate in the digital world. It researched the differences within and between participating education systems and the relationship of achievement to learning environment and student background. ICILS 2018 also included the computational thinking domain as a process of working out exactly how computers can assist people in solving problems (Fraillon et al., 2019).

The National Assessment of Educational Progress (NAEP) in the USA is one of the first large-scale online assessments in the world. President Barack Obama (2009) (https://www.cbsnews.com/news/obamas-remarks-on-education/) said that "I'm calling on our nation's governors and state education chiefs to develop standards and assessments that don't simply measure whether students can fill in a bubble on a test, but whether they possess twenty-first century skills like problem-solving and critical thinking and entrepreneurship and creativity". This reflects a trend toward the use of novel methods and techniques in assessment. The NAEP started in 1969. The largest nationwide, continuous, representative assessment in the USA, it focuses on what students know and can do in various subject areas. At the turn of the millennium, a project was designed to explore the use of technology, especially the use of the computer, as a tool to enhance the quality and efficiency of educational assessments, particularly the NAEP. In 2001, the math online (MOL) study was the first field investigation; it was followed by the writing online (WOL) project in 2002 and the problem-solving in technology-rich environments project in 2003. It investigated how CBA can be used to measure skills that cannot be measured with a PP test (Beller, 2013). In the second stage of development in 2009, almost ten years later, interactive computer tasks were administered in science. 2011 saw the launch of a CB writing assessment, with scenario-based tasks following in 2014. From 2017, the NAEP assessment was fully computerised.

Another national assessment in the USA, the Smarter Balanced Assessment Consortium (SBAC), began in 2014. It tested students using computer-adaptive technology that tailors questions to students based on their answers to previous questions. The SBAC continued to use one test at the end of the year for accountability purposes but created a series of interim tests to inform students, teachers and parents as to whether students are on track (SBAC, 2016). Table 1 summarises the year of the transition to CBA among LSA from the NAEP in 2001 to the TIMSS in 2019.

Table 1 From PP to CB: the transition year for the main LSA

LSA	*Start of transition from PP to CB*	*Transition completed*
NAEP	2001	2017
PISA	2006	2019
ICILS (started as CB)	2013	2013
SBAC (started as CB)	2014	2014
PIRLS	2016	n.d. (2021 – both versions in parallel)
TIMSS	2019	n.d.

Note: n.d.: no date is given.

5 Media comparison studies: CBA vs. PP assessments

Over the past two decades, various media studies have been carried out to determine the effect of delivery medium on students' test achievement (Oz and Ozturan, 2018). We conducted a review of these studies (see Table 2) to obtain a comprehensive overview of the main results in the Google Scholar database. As a first step, we defined the keywords,

all connected to the topic of media comparison studies. These studies evaluate the comparability issues (e.g., validity, reliability, objectivity, advantages, costs and effect on test results) of different delivery modes, that is online testing, face-to-face testing and PP testing. We used the following terms separately during a Google Scholar search: media study in CBA; PB vs. CBA; TBA/PB assessment; CBA/PB assessment; TBA/PP assessment; and comparison between PB assessment and CBA. As a second filter, we only focused on studies where the same construct was assessed in both modes, CBA and PP, and established after the turn of the millennium. Table 2 summarises these studies according to age level and sample size, field of study, country and main results.

A meta-analysis of these studies shows various results on the effect of media on students' test scores, i.e., on students' achievement. More specifically, some of these results demonstrated a significant difference between the two testing modes in favour of CB mode (e.g., Blazek and Forbey, 2011; Clariana and Wallace, 2002; Hakim, 2017; Karadeniz, 2009), while others found the opposite result of participants performing better in PP mode (e.g., Al-Amri, 2009; Csapó et al., 2009). Still other studies reported no significant differences in the two testing modes (e.g., Akdemir and Oğuz, 2008; Bodmann and Robinson, 2004; Cagiltay and Zalp-Yaman, 2013; Garas and Hassan, 2018; Hensley, 2015; Higgins et al., 2005; Horkay et al., 2006; Khoshsima and Hashemi, 2017; Logan, 2015; Mojarrad et al., 2013; Retnawati, 2015).

Beyond the actual test scores, some of the media studies also investigated participants' perceptions, attitudes and opinions with regard to the two-delivery medium. Donovan et al. (2007) explored students' opinions on the application of CBA instead of PP testing. According to the results of the survey-based study, 88.4% of the students preferred CBA to PP. Llamas-Nistal et al. (2013) confirmed this result, with 43 students out of 52 choosing online testing over traditional assessment methods. Tubaishat et al. (2006) conducted a study at university level. 59% of the students at the University of Jordan and 50% of the students at Zayed University in the United Arab Emirates liked online exams better than PP exams. Barros (2018) confirmed these findings; that is, students unequivocally preferred CB tests over PP tests.

To sum up, the differences between PP and CB test performance among secondary students and undergraduate students have been widely studied and well documented; however, there is still a gap. Very few studies have focused on the comparability issues of traditional and CB testing among kindergarten children and primary students. Most of the latest media comparison or media effect studies among secondary students have indicated that PP and CB testing are comparable and that students prefer CB tests to PP testing. Based on the few studies focusing on primary students, we can conclude that existing differences decrease over time as computers become widely accessible at schools (Csapó et al., 2014; Mayrath et al., 2012) and thus test mode effects should no longer represent an issue (Way et al., 2006), at least among secondary students.

Table 2 CBA and PP assessment of the same construct

Researchers	*Age level*	*Sample size*	*Field of study*	*Country*	*Main results*
Clariana and Wallace (2002)	Under-graduates	105	Business courses	USA	Students' achievement in the CB environment was significantly higher than that of PP mode.
Choi et al. (2003)	Under-graduates	971	English language proficiency	South Korea	A significant difference between the results in the two modes supports comparability between them.
Bodmann and Robinson (2004)	Under-graduates	55	Web-based course management system	USA	There was no significant difference.
Higgins et al. (2005)	Fourth grade	219	Reading comprehension	USA	There were no statistically significant differences.
Horkay et al. (2006)	Eighth grade	1,308	Writing assessment	USA	There were no differences in students' writing skills in the two media.
Schatz and Putz (2006)	Under-graduates	30	Management and assessment of sports-related concussion	USA	Significant but modest correlations were found between the modes.
Akdemir and Oğuz (2008)	Under-graduates	47	Educational measurement course	Turkey	Test scores were not different for the CB and PP tests.
Csapó et al. (2009)	Fifth graders (11 years old)	5,000	Mathematics and reading comprehension	Hungary	Participants' achievement was lower in CB testing than in the PP format.
Karadeniz (2009)	Under-graduates	38	Computer hardware and microprocessors course	Turkey	A significant difference was found in the scores in favour of TBA.
Al-Amri (2009)	University medical students	167	English reading tests	Saudi Arabia	Students' achievement in PP mode was significantly better than in CB.
Blazek and Forbey (2011)	Under-graduate students	387	Psychopathology test	USA	There were some significant differences in favour of CB.

Table 2 CBA and PP assessment of the same construct (continued)

Researchers	*Age level*	*Sample size*	*Field of study*	*Country*	*Main results*
Cagiltay and Zalp-Yaman (2013)	First-year engineering students	209	Chemistry course	Turkey	There was no significant performance difference between PP and CB.
Mojarrad et al. (2013)	8 to 12 years	66	Reading comprehension assessments in English as a foreign language	Iran	The quantity of reading comprehension did not differ considerably.
Csapó et al. (2014)	First-grade children	364–435	Inductive reasoning	Hungary	PP and CB tests measured pupils inductive reasoning skills very similarly, not only at the overall test level, but at the item level as well.
Logan (2015)	Grade 6	804	Mathematics	Singapore	There were no statistically significant differences.
Hensley (2015)	Grades 4–5	155	Mathematics	USA	There was no difference found in performance on PP and CB tests based on overall performance in mathematics.
Retnawati (2015)	Adults	600	Test of English proficiency	Indonesia	The reliability between the scores for the CB and PP tests was almost the same.
Khoshsima and Hashemi (2017)	Under-graduate students	228	Language knowledge and proficiency	Iran	Test-takers' scores were not different in CB and PP mode.
Hakim (2017)	Foundation year students from the English language Institute	200	English language proficiency	Saudi Arabia	There were statistically significant differences between test results in PP and CB mode, with participants in CB performing better.
Hardcastle et al. (2017)	Elementary, middle and high school	34,068	Science	USA	Performance varied with different test modes according to students' age level.
Garas and Hassan (2018)	University level	78	Financial accounting courses	United Arab Emirates	There was no statistically significant difference between the students' PP and CB scores.
Fishbein et al. (2018)	Fourth and eighth grades	16,894	Maths and science	International	There was an overall mode effect.

6 Increased effectiveness and advantages of CBA

The development, spread and accessibility of technology offer extraordinary opportunities for the improvement of educational assessment. For example, CBA facilitates highly efficient data collection and more exact, more varied testing procedures to measure more complex skills and abilities and administer more realistic, application-oriented tasks in more authentic testing environments than those of PP assessments (Beller, 2013; Bennett, 2002; Breiter et al., 2013; Bridgeman, 2010; Christakoudis et al., 2011; Csapó et al., 2012; Farcot and Latour, 2009; Kikis, 2010; Martin, 2010; Martin and Binkley, 2009; Moe, 2010; Ripley, 2010; Van Lent, 2010). Its increased effectiveness and advantages can be observed on every level of assessment:

1 *The costs of testing:* Among the benefits of late proliferation are the lower costs compared to PP assessment. The following activities are necessary for each PP testing session: item writing, proofreading, task editing and test assembly; preparation for printing and printing/copying; test delivery: packing, shipping and distribution; and data collection, collecting the tests, shipping, evaluation, coding, data recording, data cleaning, running the analysis, writing feedback and storing the tests. Each activity has its own cost implications. In the case of CBA, we do not need to print, copy, pack, ship, evaluate, code or record the data. Thus, the costs of data collection can be greatly reduced (Bennett, 2003; Choi and Tinkler, 2002; Christakoudis et al., 2011; Csapó et al., 2012; Csapó et al., 2009; Peak, 2005; Rose et al., 1999; Valenti et al., 2003; Wise and Plake, 1990). An analysis of the costs of testing showed that even two-thirds of documentation costs can be saved through CBA (Rose et al., 1999). Based on Farcot and Latour's (2009) cost analysis, the initial costs of PP testing prove to be the lowest. However, this type of testing can only remain competitive in the long run if one does not need to produce many tasks and the complexity of the tasks can be low. As the number of required tasks and their complexity increase, CBA will be a more economical and sustainable method. In sum, the costs of CBA drop significantly in the medium and long term (Bennett, 2003; Choi and Tinkler, 2002; Farcot and Latour, 2009; Kuzmina, 2010; Peak, 2005).

2 *The speed and safety of test administration and data flow:* CBA makes data processing faster and easier (Csapó et al., 2012). It is safer to maintain test-taking security with user names and passwords (Kuzmina, 2010; Marriott and Teoh, 2012). The possibility of selecting questions at random or using adaptive techniques reduces cheating, thus improving safety and providing more objectivity (Marriott and Teoh, 2012). Moreover, an adaptive test algorithm allows a more precise (lower measurement error) or less time-consuming (with the same level of measurement error) assessment of levels of knowledge, skills and abilities (Frey, 2007; Jodoin et al., 2006).

3 *The option of providing immediate feedback on completion of testing* (Becker, 2004; Csapó et al., 2014; Dikli, 2006; Mitchell et al., 2002; Valenti et al., 2003) increases the efficiency of the assessment by making it possible to measure even sudden improvement among students with diagnosed atypical development; that is, it paves the way for individualised diagnostic testing beyond the predominantly summative

approach (Kettler, 2011; Redecker and Johannessen, 2013; Van der Kleij et al., 2012).

4 *Indicators of test goodness and efficiency:* The behaviour of the tests – that is, the generalisability of the results, the validity of the construct measured, and the objectivity of data collection and evaluation – is characterised by three indicators: reliability, validity and objectivity. These are assured when the test scores, i.e. the achievement of the students, only depend on the students' level of knowledge and skills, independent of any other factors, such as the circumstances of the data collection and the harshness of the test scorer. With technology, the level of standardisation of testing conditions can be significantly boosted, thus ruling out the uncertainty of the human factor. That is, CBA promotes an increase in the indicators of test goodness (Csapó et al., 2014; Jurecka and Hartig, 2007; Marriott and Teoh, 2012; Ridgway and McCusker, 2003). We can thus achieve improved efficiency and greater measurement precision in the assessment domains already established (Csapó et al., 2014).

5 *Options for measuring new constructs:* CBA has paved the way for the development and use of new, more complex and innovative item types beyond the more traditional first-generation CB items (e.g., multiple choice; Alruwais et al., 2018). With multimedia elements, second-generation items made it possible to create more real-life problems and a more standardised testing environment (e.g., everybody listening to the same voice) than first-generation items. Finally, third-generation tests (Greiff, 2012; Greiff et al., 2012; Ripley et al., 2009), including interaction, simulations and cooperation, facilitated the measurement of construct, a feature which would be impossible with traditional assessments that rely on standard item formats (e.g., complex problem-solving (CPS); see Dörner and Funke, 2017; Greiff et al., 2012; in PISA 2012, it was called creative problem-solving). With second- and third-generation tests, we can replicate complex, real-life situations and use authentic tasks, interactions, dynamism, virtual worlds and collaboration within the test to measure even more complex, 21st century skills (Pachler et al., 2010; Ridgway et al., 2004), thus increasing the quality of educational assessment.

6 *Student motivation towards testing changes* (Meijer, 2010; Sim and Horton, 2005)*:* Technology allows creative task presentation through innovative item development opportunities (Pachler et al., 2010; Strain-Seymour et al., 2009), thus raising the motivation and enjoyment level of the assessment in a way that would have been impracticable in the PP environment. CBA can provide test environments that are similar to entertainment activities (Ridgway et al., 2004).

7 *Effective tools for logging and analysing contextual data* (e.g., time on task and number of student attempts to modify solutions; Csapó et al., 2014), *not only observed variables.* Logfile analysis, educational data mining and learning analytics offer new indicators beyond traditional test results, thus making it possible to conduct a more thorough analysis of the student's behaviour and the structure of the knowledge, skills and abilities measured.

7 Challenges and drawbacks of using TBA

Despite the many advantages TBA and CBA offer educational researchers, they also face several challenges that call for further research and also involve some drawbacks. Drawbacks of TBA can be viewed as bigger challenges for the future, thus requiring further developments and researches in the field of educational assessment.

In most cases, the basic technological solutions are already available at the student and/or school level, but – as we have seen in the situation generated by COVID 19 worldwide, even at the international level – their useful integration and application in everyday school practice are limited and require further development. This integration is strongly hindered by issues of diversity, connectivity, and lack of systematicity and compatibility.

There exists no fit for all approaches to TBA. Different assessment needs require different technological conditions, that is, the same solution cannot optimally serve every possible assessment scenario (Csapó et al., 2014). Beyond the proper infrastructure (Alruwais et al., 2018), different problems arise when TBA is used e.g., for high-stakes/low-stakes testing, large-scale/small-scale data collection, standardised/unstandardised assessment, fixed/adaptive testing, summative/formative/diagnostic assessment, using more traditional/innovative item types, replacing traditional PB assessment/launching assessment of skills related to the digital word, placing students in testing centres/in the classroom environment/at home, assessing kindergarten children/primary students/secondary students and students' familiarity/lack of familiarity with TBA. Independent of the aim, place, type and methods of assessment, validity still remains an important issue.

8 Latest developments

The latest developments in the CBA revolution in the educational context highlight two points: first, we have seen a shift from summative to formative and diagnostic assessments, which better reflect students' learning needs, facilitate understanding and provide students with immediate feedback; second, logfile analysis, educational data mining and learning analytics have contributed significantly to an understanding of the phenomenon under examination and expanded the possibilities not only from a quantitative perspective, but also from a qualitative one.

8.1 From efficient testing to personalised learning: Integrating assessment into teaching by means of technology

There is no longer any question as to whether we can develop authentic, real-life, complex, high-quality tests. At the same time, summative test results have limited usefulness with regard to personalised intervention and student-level feedback in general (Csapó and Molnár, 2019). They are often used for accountability purposes, causing negative effects in testing, such as test coaching (teaching for testing) and test score inflation (see e.g., Koretz, 2018). These effects can have a harmful influence on school climate and teacher stress (Saeki et al., 2018). However, this does not mean that testing is harmful. We must change the purpose of assessment from a rather summative to a more learning-centred, personalised approach, where testing meets the individual needs of

students through a frequent, low-stakes assessment combined with prompt and proper feedback about their level of knowledge, skills and abilities (Umami, 2018). Formative and diagnostic CB testing helps personalise learning with effective, adaptive learning and instruction programs (Grant and Basye, 2014). Teachers can use assessment platforms and programs to assess student performance before, during and after learning, which can be used to identify domains of weakness and strength and to promote directed personalised instruction (Grant and Basye, 2014). With TBA, teachers are no longer limited to standardised, yearly, summative exams or periodical, summative classroom tests. They have the opportunity to provide feedback at every step of the learning process and to use these regular assessments to measure the progress of educational objectives for individual students (Cole, 2008). Regular feedback enables teachers to tailor instruction and to aid in students' development more effectively by supplying more frequent information to parents on their children's learning progress (Grant and Basye, 2014).

TBA also makes it possible to fit the difficulty level of the tasks to the ability level of the students by giving students more difficult or less challenging questions. Through this adaptive approach, both the motivation level of the students and the information extracted during testing can be increased and the measurement error decreased.

8.2 Options in logfile analysis, educational data mining and learning analytics are increasing

Contextual information plays a significant role in educational assessment, contributes to a deeper understanding of the phenomenon under examination and can provide answers to research questions which could not be answered with traditional assessment techniques. Traditional assessment methods supply the researcher with very few indicators, such as test scores (quantitative) or subjective feedback (qualitative) from students on the testing/training session. Technology makes it possible to log, collect and analyse students' behaviour during the testing/learning session (e.g., the time needed to execute the task, the number of student attempts to adjust solutions, and the location and number of clicks made by students during the task and during the test) and thus to quantify even qualitative developmental differences to better understand the fine mechanisms of the phenomenon under examination. However, logfile data are collected more often than they are analysed (Bruckman, 2006).

Table 3 summarises the number of publications in Scopus as of 2011 with these keywords (phrases) restricted and filtered to these domains and illustrates the ever growing importance and role of logfile analysis in the social sciences and psychology, including time on task, learning analytics, educational data mining and big data. The keywords were used separately, filtered for the domains of the social sciences and psychology and resulting in 60 hits for the year. Based on the results, we can conclude that the history of the analysis of all kinds of log data dates back to 2010. In the last ten years, the number of publications focusing on an analysis of logged data has grown immensely. The most often used state-of-the-art terms are educational data mining, learning analytics and big data.

Table 3 Search results in Scopus for keywords filtered for the social sciences and psychology (6 December 2019)

Keywords	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*
Logfile or log-file analysis	4	11	8	4	9	11	13	11	11	11
Time on task	27	27	45	45	50	43	49	67	66	60
Educational data mining	8	12	24	22	38	52	68	87	88	124
Learning analytics	0	5	39	65	144	188	287	330	357	410
Big data	6	7	44	178	426	892	1,485	1,520	1,993	1,647

In the following, we only focus on papers containing the phrases ‘logfile analysis’ or ‘log-file analysis’ with the results of those papers illustrating how it is possible to use this type of analysis to quantify qualitative developmental differences to learn more about the phenomenon under examination beyond the score data. These papers also use state-of-the-art analysis (e.g., latent profile analysis) in most cases and go far beyond the possibilities of classical test theory (which is often used in time-on-task analyses). Several of them focus on students’ problem-solving behaviour on third-generation tests (e.g., Greiff et al., 2018, 2015, 2016; Herde et al., 2016), with a similarly focused paper beyond Scopus found through Google Scholar. As a result of the analyses, qualitatively different exploration strategies have been defined in a CPS environment (Greiff et al., 2018). It has been confirmed that using a theoretically effective strategy does not always result in high performance and that awareness also plays an influential role in problem-solving. The analyses have identified qualitatively different problem-solving class profiles. The most interesting group is that of rapid learners. These students start out as non-performers in their exploration behaviour in the first problem-solving scenarios but show a rapid learning curve and reach the same high level of exploration behaviour by the end of the test as proficient explorers. However, their final score is exactly the same as those who are high performers on the easiest problems, but low performers on the complex ones, with no so-called intermediate strategy users identified. Generally, the analyses have expanded the scope of previous studies and made it possible to detect a central component of children’s scientific reasoning and problem-solving behaviour.

These opportunities and research results are expected to revolutionise education. We are thus able to predict what types of activities would be most beneficial for different students, contributing significantly to the personalisation of education (Wise, 2019). According to Johnson et al. (2016), learning analytics is one of the most significant developments of the 21st century. Score-based data and analyses from previous educational research have provided opportunities for post-correction, intervention and modification (e.g., improvement and refinement of tests), with almost all of these data and analyses being output-oriented. Learning analytics enables us not only to confirm that a particular learning unit has been mastered, but also to monitor the learning activity in real time. Based on these data, both computer-controlled and human-driven techniques can be used to better tailor education to the needs of learners, thus moving away from a one-size-fits-all approach (Wise, 2019).

9 Perspectives in the present and challenges for the future

Different areas can be distinguished by discussing the perspectives on educational assessment based on the developments and experiences of the last twenty years. In our view, these developments can enhance the efficiency and efficacy of assessment, thus maximising students' engagement, motivation and learning (Adesope and Rud, 2019) if they are used not for its own sake (Gonski et al., 2018), but in an integrated and combined way that provides links between assessment, teaching and learning (Neumann et al., 2019).

Innovative technologies combine to form an integrated multi-sensory interactive application to present information to students and thus offer exciting opportunities to increase the efficiency of assessments that are more useful for teachers and more supportive, motivating and effective for students (Gonski et al., 2018; Koomen and Zoanetti, 2018). However, the real advantage of these technologies, such as touchscreens, augmented reality (AR), virtual reality (VR), mixed reality (MR), robots and behavioural monitoring (e.g., voice recognition, eye gaze, face recognition and touchless user interface) can be effectively used if they are linked to the proper assessment, and educational and developmental theories and methods. However, ways, models and theories must be devised to adapt these technologies to the human mind, including how we learn, and experimental research evidence is needed to determine which instructional features maximise learning outcomes and promote learning processes (Adesope and Rud, 2019). The systematic introduction and application of TBA in everyday school practice, including TBA, using the most common technologies (as we saw in the quarantine situation worldwide because of COVID-19) or even emerging ones, require further research and provide new challenges for educational researchers.

New learning and assessment theories and the reconceptualisation of research are needed – integrating models on multimedia learning, machine learning, learning analytics, educational data mining, knowledge representation, developmental psychology and assessment, including visualisation of the results to support human learning (Bottou, 2014; Markauskaite, 2010; Martin and Sherin, 2013; Mayer, 2009) – to maximise the use and possibilities of these tools to enhance and facilitate students' learning instead of merely summarising the current state of their knowledge based on the answer data given, which has been in the focus of educational assessment in the last 20 years. TBA can provide

1 fine-grained, process-oriented data, which can open up new possibilities to understand how we learn (Kramer and Benson, 2013)

2 knowledge which supports personalised learning with constructive feedback. The ability to use available tools calls for new assessment theories (e.g., a more detailed analysis of logfiles and process data beyond the commonly used latent profile and time-on-task analysis).

Developments in TBA are moving toward intelligent systems that facilitate students' personalised learning and monitor their emotional and cognitive status, where continuous diagnostic adaptive assessment techniques provide a challenging multimedia learning environment for the user.

The possibilities are becoming almost unlimited; however, implementing them in everyday school practice requires a great deal of research, development and time. As an

example of the very long implementation process, in the 1960s, Rasch published the Rasch model, the well-known and broadly used one-parameter item response theory model. This largely established the basis for adaptive testing, a special form of CBA that is adaptive to each test-taker's ability level. Empirical studies in the 1980s (e.g., Weiss and Kingsbury, 1984) proved that computer-adaptive testing is more effective, reduces testing time without deteriorating measurement precision and strongly increases test-takers' motivation compared to fixed tests, that is, tests comprising the same items for everybody. It took almost 40 years between demonstrating empirical evidence for the effectiveness of the Rasch model and applying it in the most prominent LSA, OECD PISA (please note that PISA was launched in 2000; that is, in the history of PISA, it took almost 20 years.)

10 Limitations

Limitations of the study include the sampling procedure. We restricted the sample to the large research databases on Google Scholar and Scopus. In other words, papers, dissertations and documents which are not indexed in Google Scholar or Scopus were excluded from the analyses. In addition, searches in Scopus were filtered further for the social sciences and psychology; that is, papers which are not indexed in these domains were also excluded from the analyses. We focused on the most prominent, mostly international LSA and excluded other research developments by analysing the effect of LSAs on TBA.

11 Discussion and conclusions

The ICT revolution has reshaped society, required new competences, and opened up new possibilities and challenges in educational assessment. Measuring and developing 21st century skills (Borodina et al., 2019) requires new assessment which goes beyond testing knowledge and provides prompt, meaningful feedback for learners and teachers as well. Traditional assessment methods are sorely lacking in this regard.

The development encompasses three main steps which lead to ever growing possibilities in educational assessment. First-generation CB tests looked very similar to traditional PP testing, but already used several advantages of CBA (e.g., feedback time and delivery mode). Second-generation CBA includes multimedia elements and makes adaptive testing possible. While employing third-generation tasks, even very complex constructs can be measured (e.g., 21st century skills) by activating interaction, simulation, cooperation and dynamically changing items. To sum up, technology plays an important role in the development of educational assessment (RQ1), and we observed a significant effect of large-scale international assessments on the evolution of TBA (RQ2).

A number of media studies were conducted around the turn of the millennium, when CBA emerged as a real alternative to PP testing even in LSA. The results were divergent because of the different samples, knowledge, skills and abilities assessed, and item formats used, but the eventual differences between PP and CB delivery mode and students' test performance have been widely studied and well documented. The latest studies have clearly indicated that PP and CB tests are comparable. Some of these results demonstrated a significant difference between the two testing modes in favour of CB

mode, while others found the opposite result. Still other studies reported no significant differences in the two testing modes. If there are differences, they decrease over time as computers become widely accessible with students preferring CB tests to PP testing. Thus, with test mode effects no longer an issue, we can concentrate on the further possibilities of the new technologies in educational assessment (RQ3).

The use of technology has greatly improved the efficiency of testing procedures: it speeds up data collection, supports real-time automatic scoring, accelerates data processing, facilitates immediate feedback and revolutionises the whole process of assessment, including innovative task presentation (for a detailed discussion of technological issues, see Csapó et al., 2012). Also, it provides new opportunities in item and test development. Beyond these options, technology makes it possible to store and analyse contextual data. This new approach is often called educational data mining, logfile analysis or learning analytics, each representing a slightly different form of analysis. Because of the many advantages, the most important assessments in the near future will probably be administered in a technological environment; however, there is still a need for further research and development on the application of CBA among kindergarten children and its systematic integration into everyday school practice (RQ4).

This trend is explicitly noticeable in the most prominent international large-scale summative assessments (e.g., IEA TIMSS and PIRLS; OECD PISA). In the last few years, taking advantage of one of the greatest possibilities of CBA, automatic feedback, there has been an emphasis on individualised diagnostic assessment beyond the mainly summative approach, thus using the power of prompt, proper feedback to personalise learning and instruction (Shatunova et al., 2019). That is, there is a need for an advanced use of the advantages and possibilities of TBA in the learning process to shift the aim of assessment from effective summative testing to personalised learning (RQ5).

Undoubtedly, CBA will replace PP at all levels of testing – summative or formative, low- or high-stakes – and offers new opportunities in assessment (e.g., online diagnostic assessment, adaptive testing, embedded assessment, measuring new constructs and learning more about students' test-taking behaviour by analysing logfiles). The technology further expands the possibilities not only from a quantitative perspective, but also from a qualitative one, thus strengthening the use of CBA (Csapó et al., 2012).

Acknowledgements

This study has been conducted with support from the National Research, Development and Innovation Fund of Hungary, financed under the OTKA K135727 funding scheme, and was also funded by EFOP-3.4.3-16-2016-00014.

References

Adesope, O.O. and Rud, A.G. (2019) 'Maximizing the affordances of contemporary technologies in education: promises and possibilities', in O.O. Adesope and A.G. Rud (Eds.): *Contemporary Technologies in Education*, pp.1–16, Springer Nature, Cham.

Akdemir, O. and Oguz, A. (2008) 'Computer-based testing: an alternative for the assessment of Turkish undergraduate students', *Computers & Education*, Vol. 51, No. 3, pp.1198–1204, https://doi.org/10.1016/j.compedu.2007.11.007.

Al-Amri, S.S. (2009) *Computer-based Testing vs Paper-based Testing: Establishing the Comparability of Reading Tests Through the Evolution of a New Comparability Model in a Saudi EFL Context*, Doctoral dissertation, The University of Essex.

Alruwais, N., Wills, G. and Wald, M. (2018) 'Advantages and challenges of using e-assessment', *International Journal of Information and Education Technology*, Vol. 8, No. 1, pp.34–37, https://doi.org/10.18178/ijiet.2018.8.1.1008.

American Psychological Association (APA) (1986), *Guidelines for Computer-based Tests and Interpretations*, Committee on Professional Standards, American Psychological Association, Board of Scientific Affairs, Committee on Psychological Tests, & Assessment, The American Psychological Association.

Baker, E.L. and Mayer, R.E. (1999) 'Computer-based assessment of problem solving', *Computers in Human Behavior*, Vol. 15, Nos. 3–4, pp.269–282.

Barros, J.P. (2018, March) 'Students' perceptions of paper-based vs. computer-based testing in an introductory programming course', *10th International Conference on Computer Supported Education, CSEDU 2018*, Vol. 2, pp.303–308, SciTePress, https://doi.org/10.5220/0006794203030308.

Becker, J. (2004) *Computergestütztes Adaptives Testen (CAT) von Angst entwickelt auf der Grundlage der Item Response Theorie (IRT)*, Unpublished PhD dissertation, Freie Universität, Berlin.

Beller, M. (2013) 'Technologies in large-scale assessments: new directions, challenges, and opportunities', in M. von Davier, E. Gonzalez, I. Kirsch and K. Yamamoto (Eds.): *The Role of International Large-scale Assessments: Perspectives from Technology, Economy, and Educational Research*, pp.25–45, Springer, Dordrecht, Netherlands.

Bennett, R.E. (2002) 'Using electronic assessment to measure student performance: online testing', *State Education Standard*, Vol. 3, No. 3, pp.23–29.

Bennett, R.E. (2003) *Online Assessment and the Comparability of Score Meaning*, Educational Testing Service, Princeton, NJ.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M. and Rumble, M. (2012) *Defining Twenty-first Century Skills*. *In Assessment and Teaching of 21st Century Skills*, pp.17–66, Springer, Dordrecht.

Blazek, N.L. and Forbey, J.D. (2011) 'A comparison of validity rates between paper-and-pencil and computerized testing with the MMPI-2', *Assessment*, Vol. 18, No. 1, pp.63–66, https://doi.org/10.1177/1073191110381718.

Bodmann, S.M. and Robinson, D.H. (2004) 'Speed and performance differences among computer-based and paper-pencil tests', *Journal of Educational Computing Research*, Vol. 31, No. 1, pp.51–60, https://doi.org/10.2190/GRQQ-YT0F-7LKB-F033.

Borodina, T., Sibgatullina, A. and Gizatullina, A. (2019) 'Developing creative tfhinking in future teachers as a topical issue of higher education', *Journal of Social Studies Education Research*, Vol. 10, No. 4, pp.226–245.

Bottou, L. (2014) 'From machine learning to machine reasoning: an essay', *Machine Learning*, Vol. 94, No. 2, pp.133–149.

Breiter, A., Groß, L.M. and Stauke, E. (2013) 'Computer-based large-scale assessments in Germany', in D. Passey, A. Breiter and A. Visscher (Eds.): *Next Generation of Information Technology in Educational Management*, pp.41–54, Springer, Berlin, Heidelberg.

Bridgeman, B. (2010) 'Experiences from large-scale computer-based testing in the USA', in F. Scheuermann and J. Bjornsson (Eds.): *The Transition to Computer-based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing*, pp.39–44, European Communities, Brussels.

Bruckman, A. (2006) 'Analysis of log file data to understand behavior and learning in an online community', in J. Weiss, J. Nolan, J. Hunsinger and P. Trifonas (Eds.): *The International Handbook of Virtual Learning Environments*, pp.1449–1465, Springer, Dordrecht, Netherlands.

Cagiltay, N. and Ozalp-Yaman, S. (2013) 'How can we get benefits of computer-based testing in engineering education', *Computer Applications in Engineering Education*, Vol. 21, No. 2, pp.287–293.

Choi, I-C., Kim, K.S. and Boo, J. (2003) 'Comparability of a paper-based language test and a computer-based language test', *Language Testing*, Vol. 20, No. 3, pp.295–20, https://doi.org/10.1191/0265532203lt258oa.

Choi, S.W. and Tinkler, T. (2002) 'Evaluating comparability of paper and computer based assessment in a K-12 setting', paper presented at the *Annual Meeting of the National Council on Measurement in Education*, New Orleans, LA.

Christakoudis, C., Androulakis, G.S. and Zagouras, C. (2011) 'Prepare items for large scale computer based assessment: Case study for teachers' certification on basic computer skills', *Procedia-Social and Behavioral Sciences*, Vol. 29, pp.1189–1198.

Clariana, R. and Wallace, P. (2002) 'Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, Vol. 33, No. 5, pp.593–602, https://doi.org/10.1111/1467-8535.00294.

Cole, R.W. (2008) *Educating Everybody's Children: Diverse Teaching Strategies for Diverse Learners*, ASCD.

Csapó, B. and Molnár, G. (2019) 'Online diagnostic assessment in support of personalized teaching and learning: the eDia system', *Frontiers in Psychology*, Vol. 10, p.1522, https://doi.org/10.3389/fpsyg.2019.01522.

Csapó, B., Ainley, J., Bennett, R.E., Latour, T. and Law, N. (2012) 'Technological issues for computer-based assessment', in P. Griffin, B. McGaw and E. Care (Eds.): *Assessment and Teaching of 21st Century* Skills, pp.143–230, Springer, New York, https://doi.org/10.1007/978-94-007-2324-5_4.

Csapó, B., Molnár, G. and Nagy, J. (2014) 'Computer-based assessment of school-readiness and reasoning skills', *Journal of Educational Psychology*, Vol. 106, No. 2, pp.639–650.

Csapó, B., Molnár, G. and Tóth, K. (2009) 'Comparing paper-and-pencil and online assessment of reasoning skills: a pilot study for introducing TAO in large-scale assessment in Hungary', in F. Scheuermann and J. Björnsson (Eds.): *The Transition to Computer-based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing*, pp.113–118, Office for Official Publications of the European Communities, Luxemburg.

Dikli, S. (2006) 'An overview of automated scoring of essays', *The Journal of Technology, Learning and Assessment*, Vol. 5, No. 1, pp.1–30.

Donovan, J., Mader, C. and Shinsky, J. (2007) 'Online vs. traditional course evaluation formats: student perceptions', *Journal of Interactive Online Learning*, Vol. 6, No. 3, pp.158–180.

Dörner, D. and Funke, J. (2017) 'Complex problem solving: what it is and what it is not', *Frontiers in Psychology*, Vol. 8, p.1153.

Farcot, M. and Latour, T. (2009) 'Transitioning to computer-based assessments: a question of costs', in F. Scheuermann and J. Bjornsson (Eds.): *The Transition to Computer-based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing*, pp.108–16, European Communities, Brussels.

Fishbein, B., Martin, M. O., Mullis, I.V. and 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*, Vol. 6, No. 1, p.11.

Fraillon, J., Ainley, J., Schulz, W., Duckworth, D. and Friedman, T. (2019) *IEA International Computer and Information Literacy Study 2018 Assessment Framework*, Springer.

Frey, A. (2007) 'Adaptives testen', in H. Moosbrugger and A. Kelava (Eds.): *Testtheorie und Testkonstruktion*, pp.261–278, Springer, Berlin, Heidelberg.

Fulcher, G. (2000) 'Computers in language testing', in P. Brett and G. Moterram (Eds.): *A Special Interest in Computers: Learning and Teaching with Information and Communications Technologies*, pp.93–107, IATEFL Publications, Manchester.

Garas, S. and Hassan, M. (2018) 'Student performance on computer-based tests versus paper-based tests in introductory financial accounting: UAE evidence', *Academy of Accounting and Financial Studies Journal*, Vol. 22, No. 2, pp.1–14

Gonski, D. et al. (2018) *Through Growth to Achievement*, Report of the Review to Achieve Educational Excellence in Australian Schools, Australian Government [online] https://docs.education.gov.au/system/files/doc/other/662684_tgta_accessible_final_0.pdf (accessed 29 May 2021).

Grant, P. and Basye, D. (2014) *Personalized Learning: A Guide for Engaging Students with Technology*, International Society for Technology in Education.

Greiff, S. (2012) 'Assessment and theory in complex problem solving: a continuing contradiction', *Journal of Educatio nal and Developmental Psychology*, Vol. 2, No. 1, pp.49–56.

Greiff, S., Krkovic, K. and Hautamäki, J. (2015) 'The prediction of problem-solving assessed via microworlds', *European Journal of Psychological Assessment*, Vol. 32, No. 4, pp.298–306.

Greiff, S., Molnár, G., Martin, R., Zimmermann, J. and Csapó, B. (2018) 'Students' exploration strategies in computer-simulated complex problem environments: a latent class approach', *Computers & Education*, Vol. 126, pp.248–263.

Greiff, S., Niepel, C., Scherer, R. and Martin, R. (2016) 'Understanding students' performance in a computer-based assessment of complex problem solving: an analysis of behavioral data from computer-generated log files', *Computers in Human Behavior*, Vol. 61, pp.36–46.

Greiff, S., Wüstenberg, S. and Funke, J. (2012) 'Dynamic problem solving: a new assessment perspective', *Applied Psychological Measurement*, Vol. 36, No. 3, pp.189–213, https://doi.org/10.1177/0146621612439620.

Griffin, P., Care, E. and McGaw, B. (2012) 'The changing role of education and schools', in *Assessment and Teaching of 21st Century Skills*, pp.1–15, Springer, Dordrecht, Netherlands.

Hakim, B.M. (2017) 'Comparative study on validity of paper-based test and computer-based test in the context of educational and psychological assessment among Arab students', *International Journal of English Linguistics*, Vol. 8, No. 2, pp.85–91, http://doi.org/10.5539/ijel.v8n2p85.

Hardcastle, J., Herrmann-Abell, C.F. and DeBoer, G.E. (2017) 'Comparing student performance on paper-and-pencil and computer-based tests', *Grantee Submission*, San Antonio, TX.

Hensley, K.K. (2015) *Examining the Effects of Paper-based and Computer-based Modes of Assessment on Mathematics Curriculum-based Measurement*, PhD (Doctor of Philosophy) thesis, University of Iowa, https://doi.org/10.17077/etd.ireseh1q.

Herde, C.N., Wüstenberg, S. and Greiff, S. (2016) 'Assessment of complex problem solving: what we know and what we don't know', *Applied Measurement in Education*, Vol. 29, No. 4, pp.265–277.

Higgins, J., Russell, M. and Hoffmann, T. (2005) 'Examining the effect of computer-based passage presentation of reading test performance', *The Journal of Technology, Learning and Assessment*, Vol. 3, No. 4, pp.1–35.

Horkay, N., Bennett, R.E., Allen, N., Kaplan, B.A. and Yan, F. (2006) 'Does it matter if I take my writing test on computer?: an empirical study of mode effects in NAEP', *The Journal of Technology, Learning and Assessment*, Vol. 5, No. 2, Retrieved 23 November 2019 [online] https://ejournals.bc.edu/index.php/jtla/article/view/1641.

Jodoin, M., Zenisky, A. and Hambleton, R.K. (2006) 'Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes', *Applied Measurement in Education*, Vol. 19, No. 3, pp.203–220.

Johnson, L., Becker, S.A., Cummins, M., Estrada, V., Freeman, A. and Hall, C. (2016) *NMC Horizon Report: 2016 Higher Education Edition*, pp.1–50, The New Media Consortium.

Jurecka, A. and Hartig, J. (2007) 'Computer- und Netzbasiertes assessment', in J. Hartig and E. Klieme (Eds.): *Möglichkeiten und Voraussetzungen Technologiebasierter Kompetenzdiagnostik*, pp.37–48, Bundesministerium für Bildung und Forschung, Berlin, Bonn.

Karadeniz, S. (2009) 'The impacts of paper, web and mobile based assessment on students' achievement and perceptions', *Scientific Research and Essay*, Vol. 4, No. 10, pp.984–991, Retrieved November 18, 2019 [online] http://www.academicjournals.org/app/webroot/article/article1380547300_Karadeniz.pdf.

Kettler, R.J. (2011) 'Computer-based screening for the new modified alternate assessment', *Journal of Psychoeducational Assessment*, Vol. 29, No. 1, pp.3–13.

Khoshsima, H. and Hashemi Toroujeni, S.M. (2017) 'Comparability of computer-based testing and paper-based testing: testing mode effect, testing mode order, computer attitudes and testing mode preference', *International Journal of Computer (IJC)*, Vol. 24, No. 1, pp.80–99.

Kikis, K. (2010) 'Reflections on paper-and-pencil tests to eAssessments: narrow and broadband paths to 21st century challenges', in F. Scheuermann and J. Bjornsson (Eds.): *The Transition to Computer-based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing*, pp.99–103, Brussels: European Communities.

Kingston, N.M. (2008) 'Comparability of computer- and paper-administered multiple-choice tests for K–12 populations: a synthesis', *Applied Measurement in Education*, Vol. 22, No. 1, pp.22–37, https://doi.org/10.1080/08957340802558326.

Koomen, M. and Zoanetti, N. (2018) 'Strategic planning tools for large-scale technology-based assessments', *Assessment in Education: Principles, Policy & Practice*, Vol. 25, No. 2, pp.200–223, doi: 10.1080/0969594X.2016.1173013.

Koretz, D. (2018) 'Moving beyond the failure of test-based accountability', *American Educator*, Vol. 41, No. 4, pp.22–26.

Kramer, S. and Benson, S. (2013) 'Changing faculty use of technology – one cohort at a time', *Journal of Applied Research in Higher Education*, Vol. 5, No. 2, pp.202–221, https://doi.org/10.1108/JARHE-11-2012-0036.

Kuzmina, I.P. (2010) 'Computer-based testing: advantages and disadvantages', *Bulletin of the National Technical University of Ukraine Kyiv Polytechnic Institute. Philosophy. Psychology. Pedagogy* (*Вісник Національного технічного університету України Київський політехнічний інститут. Філософія. Психологія. Педагогіка*), No. 1, pp.192–196.

Llamas-Nistal, M., Fernández-Iglesias, M.J., González-Tato, J. and Mikic-Fonte, F.A. (2013) 'Blended e-assessment: migrating classical exams to the digital world', *Computers & Education*, Vol. 62, pp.72–87.

Logan, T. (2015) 'The influence of test mode and visuospatial ability on mathematics assessment performance', *Mathematics Education Research Journal*, Vol. 27, No. 4, pp.423–441, https://doi.org/10.1007/s13394-015-0143-1.

Markauskaite, L. (2010) 'Digital media, technologies and scholarship: some shapes of eResearch in educational inquiry', *The Australian Educational Researcher*, Vol. 37, No. 4, pp.79–101.

Marriott, P. and Teoh, L. (2012) *ICT for Assessment and Feedback on Undergraduate Accounting Modules*, The Higher Education Academy [online] http://www.heacademy.ac.uk/resources/detail/disciplines/finance-and-accounting/using-ICT-in-assessment-and-feedback (accessed May 2019).

Martin, R. (2010) 'Utilising the potential of computer delivered surveys in assessing scientific Literacy', in F. Scheuermann and J. Bjornsson (Eds.): *The Transition to Computer-based Assessment: New Approaches to Skills Assessment and Implications for Large-scale testing*, pp.172–177, European Communities, Brussels.

Martin, R. and Binkley, M. (2009) 'Gender differences in cognitive tests: a consequence of gender-dependent preferences for specific information presentation formats?', in F. Scheuermann and J. Bjórnsson (Eds.): *The Transition to Computer-based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing*, pp.75–82, Office for Official Publications of the European Communities, Luxembourg.

Martin, T. and Sherin, B. (2013) 'Learning analytics and computational techniques for detecting and evaluating patterns in learning: an introduction to the special issue', *Journal of the Learning Sciences*, Vol. 22, No. 4, pp.511–520.

Mayer, R.E. (2009) *Multimedia Learning*, 2nd ed., Cambridge University Press, New York.

Mayrath, M., Clarke-Midura, J. and Robinson, D. (2012) 'Introduction to technology-based assessments for 21st century skills', *Technology-based Assessments for 21st Century Skills*, pp.1–11.

Meijer, R. (2010) 'Transition to computer-based assessment: motivations and considerations', in Scheuermann, F. and Bjornsson, J. (Eds.): *The Transition to Computer-Based Assessment: New Approaches to Skills Assessment and Implications for Large-Scale Testing*, pp.104–107, European Communities, Brussels.

Mitchell, T., Russel, T., Broomhead, P. and Aldridge, N. (2002) 'Towards robust computerized marking of free-text responses', in M. Danson (Ed.): *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughboroug University, Loughborouh [online] https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/1884/1/Mitchell_t1.pdf (accessed 29 May 2021).

Moe, E. (2010) 'Introducing large-scale computerized assessment – Lessons learned and future challenges', in F. Scheuermann and J. Bjórnsson (Eds.): *The Transition to Computer-based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing*, pp.51–56, Office for Official Publications of the European, Luxembourg.

Mojarrad, H., Hemmati, F., Jafari Gohar, M. and Sadeghi, A. (2013) 'Computer-based assessment (CBA) vs. paper/pencil-based assessment (PPBA): an investigation into the performance and attitude of Iranian EFL learners' reading comprehension', *International Journal of Language Learning and Applied Linguistics World*, Vol. 4, No. 4, pp.418–428.

Mullis, I.V. and Martin, M.O. (2017) *TIMSS 2019 Assessment Frameworks*, International Association for the Evaluation of Educational Achievement, The Netherlands, Amsterdam.

Mullis, I.V., Martin, M.O., Foy, P. and Hooper, M. (2017) 'ePIRLS 2016: international results in online informational reading', *International Association for the Evaluation of Educational Achievement*.

Neumann, M.M., Anthony, J.L., Erazo, N.A. and Neumann, D.L. (2019) 'Assessment and technology: mapping future directions in the early childhood classroom', *Frontiers in Education*, Vol. 4, No. 116, doi: 10.3389/feduc.2019.00116.

Organisation for Economic Co-operation and Development (OECD) (2010) *PISA 2012 Field Trial Problem Solving Framework*, Paris.

Organisation for Economic Co-operation and Development (OECD) (2011) *Education at a Glance 2011: OECD Indicators*, p.497, Paris.

Organisation for Economic Co-operation and Development (OECD) (2013) *PISA 2015 Draft Collaborative Problem Solving Framework*, OECD Publishing, Paris.

Organisation for Economic Co-operation and Development (OECD) (2016) 'Measuring student knowledge and skills', *The PISA 2000 Assessment of Reading, Mathematical and Scientific Literacy*, Paris, https://doi.org/10.1787/9789264255425-en.

Organisation for Economic Co-operation and Development (OECD) (2014) *PISA 2012 Results: Creative Problem Solving: Students' Skills in Tackling real-life Problems*, Vol. V, Paris, https://doi.org/10.1787/9789264208070-5-en.

Oz, H. and Ozturan, T. (2018) 'Computer-based and paper-based testing: does the test administration mode influence the reliability and validity of achievement tests?', *Journal of Language and Linguistic Studies*, Vol. 14, No. 1, pp.67.

Pachler, N., Daly, C., Mor, Y. and Mellar, H. (2010) 'Formative e-assessment: practitioner cases. *Computers & Education*, Vol. 54, No. 3, pp.715–721, https://doi.org/10.1016/j.compedu. 2009.09.032.

Peak, P. (2005) *Recent Trends in Comparability Studies*, Pearson Educational Measurement, [online] http://www.pearsonassessments.com/NR/rdonlyres/5FC04F5A-E79D-45FE-8484-07AACAE2DA75/0/TrendsCompStudies_rr0505.pdf (accessed 29 May 2021).

Quansah, F. (2018) 'Traditional or performance assessment: what is the right way in assessing learners', *Research on Humanities and Social Sciences*, Vol. 8, No. 1, pp.21–24.

Redecker, C. and Johannessen, Ø. (2013) 'Changing assessment – towards a new assessment paradigm using ICT',*European Journal of Education*, Vol. 48, No. 1, pp.79–96.

Redecker, C., Ala-Mutka, K. and Punie, Y. (2010) *Learning 2.0: The Impact of Social Media on Learning in Europe*, Policy Brief, JRC Scientific and Technical Report, EUR JRC56958 EN, [online] http://bit.ly/cljlpq.

Retnawati, H. (2015) 'The comparison of accuracy scores on the paper and pencil testing vs. computer-based testing', *Turkish Online Journal of Educational Technology–TOJET*, Vol. 14, No. 4, pp.135–142.

Ridgway, J. and McCusker, S. (2003) 'Using computers to assess new educational goals', *Assessment in Education*, Vol. 10, No. 3, pp.309–328.

Ridgway, J., McCusker, S. and Pead, D. (2004) *Literature Review of E-Assessment*, hal-00190440 [online] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.189.5286&rep=rep1&type=pdf (accessed 29 May 2021).

Ripley, M. (2010) 'Transformational computer-based testing', in F. Scheuermann and J. Bjórnsson (Eds.): *The Transition to Computer-based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing*, pp.92–98, Office for Official Publications of the European Communities, Luxembourg.

Ripley, M., Harding, R., Redif, H., Ridgway, J. and Tafler, J. (2009) *Review of Advanced e-Assessment Techniques (RAeAT) Final Report*, Joint Information Systems Committee.

Rose, M., Hess, V., Hörhold, M., Brähler, E. and Klapp, B.F. (1999) 'Mobile computergestützte psychometrische Diagnostik. Ökonomische Vorteile und Ergebnisse zur Teststabilität', *Psychotherapie Psychosomatik Medizinische Psychologie*, Vol. 49, pp.202–207.

Saeki, E., Segool, N., Pendergast, L. and von der Embse, N. (2018) 'The influence of test-based accountability policies on early elementary teachers: school climate, environmental stress, and teacher stress', *Psychology in the Schools*, Vol. 55, No. 4, pp.391–403.

Schatz, P. and Putz, B.O. (2006) 'Cross-validation of measures used for computer-based assessment of concussion', *Applied Neuropsychology*, Vol. 13, No. 3, pp.151–159, DOI: 10.1207/s15324826an1303_2.

Shatunova, O., Anisimova, T., Sabirova, F. and Kalimullina, O. (2019) 'STEAM as an innovative educational technology', *Journal of Social Studies Education Research*, Vol. 10, No. 2, pp.131–144.

Sim, G. and Horton, M. (2005) 'Performance and attitude of children in computer based versus paper based testing', in Kommers, P. and Richards, G. (Eds.): *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, VA: AACE, Chesapeake pp.3610–3614.

Skinner, B.F. (1958) 'Teaching machines', *Science*, Vol. 128, No. 3330, pp.969–977.

Smarter Balanced Assessment Consortium (SBAC) (2016) *Smarter Balanced Assessment Consortium: 2014–15*, Technical Report, Los Angeles.

Strain-Seymour, E., Way, W. and Dolan, R.P. (2009) *Strategies and Processes for Developing Innovative Items in Large-scale Assessments*, Research Report, Pearson Education, Iowa City, IA.

Tubaishat, A., Bhatti, A. and El-Qawasmeh, E. (2006) 'ICT experiences in two different Middle Eastern universities', *Issues in Informing Science & Information Rechnology*, Vol. 3, pp.667–678, https://doi.org/10.28945/922.

Umami, I. (2018) 'Moderating influence of curriculum, pedagogy, and assessment practices on learning outcomes in Indonesian secondary education', *Journal of Social Studies Education Research*, Vol. 9, No. 1, pp.60–75.

Valenti, S., Neri, F. and Cucchiarelli, A. (2003) 'An overview of current research on automated essay grading', *Journal of Information Technology Education: Research*, Vol. 2, No. 1, pp.319–330.

Van der Kleij, F.M., Eggen, T.J., Timmers, C.F. and Veldkamp, B.P. (2012) 'Effects of feedback in a computer-based assessment for learning', *Computers & Education*, Vol. 58, No. 1, pp.263–272.

Van Lent, G. (2010) 'Risks and benefits of CBT versus PBT in high-stakes testing', in F. Scheuermann and J. Bjornsson (Eds.): *The Transition to Computer-based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing*, pp.83–91, European Communities, Brussels.

Wang, S., Jiao, H., Young, M.J., Brooks, T. E. and Olson, J. (2008) 'Comparability of computer-based and paper-and-pencil testing in K-12 assessment: a meta-analysis of testing mode effects', *Educational and Psychological Measurement*, Vol. 68, No. 1, pp.5–24, https://doi.org/10.1177/0013164407305592.

Way, W.D., Davis, L.L. and Fitzpatrick, S. (2006) 'Score comparability of online and paper administrations of the Texas assessment of knowledge and skills', *Annual Meeting of the National Council on Measurement in Education*, San Francisco, CA.

Weiss, D.J. and Kingsbury, G. (1984) 'Application of computerized adaptive testing to educational problems. *Journal Educational Measurement*, Vol. 21, No. 4, pp.361–375, https://doi.org/10.1111/j.1745-3984.1984.tb01040.x. (Google Scholar).

Wise, A.F. (2019) 'Learning analytics: using data-informed decision-making to improve teaching and learning', in O. Adesope and A.G. Rudd (Eds.): *Contemporary Technologies in Education: Maximizing Student Engagement, Motivation, and Learning*, pp.119–143, Palgrave Macmillan, New York.

Wise, S.L. and Plake, B.S. (1990) 'Computer-based testing in higher education', *Measurement and Evaluation in Counseling and Development*, Vol. 23, No. 1, pp.3–10.