DIGITALIZED INTERACTIVE ITEM COMPONENTS IN COMPUTER-BASED-ASSESSMENT IN MATHEMATICS FOR K12 STUDENTS: A RESEARCH SYNTHESIS

Moosa A. A. Alhadi, EdM
Rutgers University, NJ, USA
maa405@scarletmail.rutgers.edu

Dake Zhang, PhD.
Rutgers University, NJ, USA
dake.zhang@gse.rutgers.edu

Ting Wang, PhD.
ETS, NJ, USA
twang001@ets.org

Carolyn A. Maher, Ed.D.
Rutgers University, NJ, USA
carolyn.maher@gse.rutgers.edu

This research synthesizes studies that used a Digitalized Interactive Component (DIC) to assess K-12 student mathematics performance during Computer-based-Assessments (CBAs) in mathematics. A systematic search identified ten studies that categorized existing DICs according to the tools that provided language assistance to students and tools that supported students problem solving. We report on the one study that involved students with learning disabilities and three studies involved English Language Learners. One study focused on assessing geometry content and four studies targeted on number and operations understanding. For other studies included a mixture of mathematics domains. Mixed results were reported as to the effectiveness of the availability of DICs. The research suggests that older children were more likely to benefit from availability of the DIC than younger children, and that DICs have greater impact on students with special needs.

Keywords: Computerized mathematics tests, mathematics assessment, Interactive digitalized components

While there are multiple literature reviews evaluating the effectiveness of Computer-based-Assessment (CBA), no systematic review was found in evaluating the effects of Digitalized Interactive Components (DICs) that are used in CBAs. Computer- based-assessments (CBA) have been widely used in national and international large-scale tests as part of recent reform efforts stressing teaching, learning and testing in digital contexts. Important CBAs include high-stake state standard assessments (e.g., the Partnership for Assessment of Readiness for College and Careers, PARCC, the Smarter Balanced Assessment), the national report card (National Assessment of Educational Progress, NAEP), and the international comparison assessments (e.g., Program for International Student Assessment, PISA). The shift towards implementing CBAs for assessment can be an important improvement during testing, potentially impacting students' engagement, interpretations, and responses to test items (Jerrim, 2016).

CBAs are created to optimize learning goals and assessment techniques through high-quality tests (Smoline, 2008), and to improve the participation and performance of students with disabilities (Flowers et al., 2011). Computers provide the opportunity to offer tasks that display animated or dynamic changes in different aspects over time: Interactive games, simulations, and microworlds, that can be made by computers, enable exploring problems, discovering rules, forming relationships, and developing successful strategies. Computers also can help students deal with complicated data sets (Ridgway & McCusker, 2003). In a study to compare CBAs with pen-and-paper assessments (PPAs) for students with a read-aloud accommodation, Flowers et al. (2011) found that students prefer CBAs and report that they perform better when CBAs are available. According to Hoogland and Tout (2018), CBAs have advantages over traditional paper and pencil assessments (PPAs) in: Supporting the assessment of high-level mathematical thinking process, illustrating authentic problems from our real-life settings to implement mathematical concepts, and building tests and analyzing results through complicated psychometric processes with new techniques in scoring and reporting.

Digital-Interactive-Components

As a key element of CBA, DICs are part of the item stem with which students can interact and employ to answer the question. For example, an interactive ruler is an on-screen ruler that students can use to measure the length of objects in the screen. DICs commonly used in the current CBAs include, but are not limited to, drag-and-drop, annotations, changing the color of a shape upon clicking, etc. The NAEP items that have access to the touch-screen tablets mode include DICs such as using a read-aloud tool, making digital notes, adjusting screen themes, highlighting texts, dynamic texts, interactive maps, embedded digital calculators, interactive graphs, and virtual simulations (National Center for Educational Statistics, NCES, n.d.).

DICs enable practitioners to measure students' knowledge and skills in ways more authentic to a CBA environment (Bergner & von Davie, 2019; Lissitz & Jiao, 2012; Masters & Gushta, 2018), and to improve the measurement of student learning by targeting constructs that are otherwise difficult to capture with the PPA mode (Bennett et al., 2008; Chen & Perie, 2018; Jerrim, 2016). DICs are expected to offer benefits to make CBAs over traditional PPAs by engaging students in simulated real-world tasks and measuring aspects of a construct that cannot be measured with PPA items (Huff & Sireci, 2001). Some interactive tools require students to work with them in the process of producing products/responses, a more authentic form of measurement (Archbald & Newmann, 1988; Bennett, 1999; Harlen & Crick, 2003; Huff & Sireci, 2001; Masters & Gushta, 2018; McFarlane, et al., 2000). Interactive features of DICs have the potential to improve assessment accessibility for some disadvantaged groups, such as students with disabilities and English Learners (ELs). Further, technologies involving the use of DIC tools can enable the collection of real-time process data that could offer diagnostic assessment information for educators to support improved teaching and student learning (Jiang et al., 2021). Despite the potential benefits of DIC inclusion in CBAs and the emerging trend of their inclusion in testing, implementations of DICs in testing have been challenging.

A first challenge in current assessment practices is the lower percent of DICs in all CBA items. Despite the fast-growing research on educational technology (Lindner, 2020), many of the studies in the literature focus on technology as a tool for instruction, whereas sparse research has been conducted on technology as a tool for assessment (Cheng & Basu, 2009). Moreover, even with knowledge from the current CBA literature, computerized test items often apply traditional formatting similar to PPAs, such as multiple choice, true-and-false, and fill-in-the-blanks (Cheng & Basu, 2009), whereas the use of interactive features in rarely reported (Dindar et al., 2015).

A second challenge in the current assessment implementation of DICs lies in the poor design of existing DICs. Invalidated DICs might introduce construct-irrelevant variance even if students have mastered the content domain, thus diminishing validity and utility of scores for students of certain subgroups. Russel and Moncaleano (2019) examined the use and construct fidelity of technology-enhanced items employed in international K12 large-scale assessment programs, and reported that only 40% of the technology-enhanced items has good fidelity in measuring the aimed construct. In particular, a sizable percentage of drag-and-drop items has a low level of fidelity. Likewise, student cognitive interviews in NAEP 2016 indicated that students experienced tremendous confusions when interacting with DICs. More than half of the students encountered various difficulties using DICs, even though more than half of the students had experiences with similar tools. Their confusion centered around figuring out how to use the tools to respond to the assessment items rather than applying the tools in their problem solving.

Even more alarmingly, poorly designed DIC tools can become a source of bias towards students from under-represented groups. Studies show that students need to be trained to use computerized items that involve DICs since the absence of learning to use the tools properly can negatively impact potential benefits of DICs (Olson, 2005). Students with less access to learn digital skills are more vulnerable to benefit with potentially negative impacts on their performance using poorly designed DIC features (Jerrim, 2016; Schatz, et al., 2010; Stickney et al., 2012). The extent to which a disparity might be present for some students is still largely unexplored (Wang et al., 2021). According to Buzick, (2021), with the exception of embedded calculators, students with disabilities showed an even lower frequency of using any DICs than their peers without disabilities, even though all embedded DICs were rarely used in all students. It still remains unclear what roles specific types and features of DICs play in measuring students' knowledge and skills, and how these features might impact the performance of student subgroups. For example, research to date has yet to examine the degree to which items with DIC yield construct irrelevant variance for different subgroups of students with different socioeconomic status (SES) and gender.

Although there have been systematical reviews pertinent to comparing the different effects of PPA and CBA for K12 students on mathematics related measures, there is an absence of research reviews specifically on the availability and effects of existing DICs. To this end, we offer a research synthesis with the goal of exploring the use of interactive tools from research studies investigating students' mathematics performance. Questions that guide our review are:

1. What existing DICs are available for use in assessing student learning mathematics in the literature of CBA in mathematics? Specifically, what are the main features and purpose of the tools? How effective are they in capturing student performance? What student characteristics, such as age, gender, ability level, influence the effectiveness of these interactive tools?
2. What mathematics domains are involved in the reviewed DIC studies? Are the DICs more effective within certain mathematical content domains? Are certain item features more or less relevant?

Method

Search Procedures

For this literature review, we followed guidelines from the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA; Moher et al. 2009) and three approaches were completed to accurately search for peer-reviewed journal articles. We conducted (a) an electronic search of databases with key words related to computer-based assessment, (b) another round of electronic search of databases with key words of specific DIC types, and (c) an ancestral search to identify any additional relevant articles.

First, we searched five databases: Academic Search Premier, ERIC, ProQuest Education Database, Teacher Reference Center, and APA PsycInfo (including PsycArticles). Within the abstracts of English studies, we searched using a combination of the following terms: "Computer based, assessment, mathematic*" and "Computer based, Test*, mathematic*" The total of the resulted studies was 1589, which were screened to 39 potentially viable studies for inclusion, based on titles and abstracts. Full-text review of each study led to an inclusion of eight articles for this synthesis.

Secondly, we also searched additional key words emphasizing interactive features with the above databases, with combined terms of "interactive technology", "computer* feature", "digit* feature", "interact* computer*", as well as specific DICs from the studies identified from the first round of search, including "calculator" "glossary" "drag drop" "animat*" and "digit* ruler." Every single phrase was attached using AND condition with the two terms "assessment" and "mathematic*." The total of the resulted studies was 444 and 2 additional studies were identified.

Thirdly, we examined the included studies reviewed by related meta-analysis or review synthesis. Specifically, we reviewed mathematics studies included in the three prior meta-analyses that was conducted by Kingston (2009), Wang et al. (2007), and by Kingston (2009) that reviewed research comparing PPA vs. CBA effects. No additional studies were identified beside these three meta-analysis studies.

Inclusion and Exclusion Criteria

In the research synthesis we only included quantitative studies that investigated the effectiveness of using DICs in assessing mathematics performance for K-12 students. Studies included in the review were required to meet all the three criteria: (a) inclusion of quantitative data of students' mathematics assessment results in CBA with specific DICs; (b) all participants were K-12 students; and (c) studies were published in peer-reviewed journals in English. Ten studies were identified, including six studies comparing a DIC condition with a non-DIC condition, and four studies exploring specific features of DIC conditions.

The following types of studies were excluded: (a) those that provided feedback or corrections in an assessment (e.g., Hippisley et al., 2005); (b) those based on computer-based programs with DICs as interventions rather than assessments (e.g., Crawford et al., 2016; Gambari et al., 2014; Ninness et al., 1998); (c) unpublished dissertations, conference presentations, technical reports (e.g., Bridgeman & Potenza, 1998; Christensen et al., 2014; Price et al., 2014); (d) qualitative studies (Pantzare, 2012); (e) focusing on non-mathematics subjects (e.g., Koong & Wu, 2010 on social studies; Cheng & Basu, 2009 on science); (f) those with participants not in K-12 (e.g., Arslan et al., 2020; Gallagher et al., 2002; Gulley et al., 2017; Mavridis & Tsiatsos, 2017), or (g) no specific DICs described (e.g., Logan, 2015).

Coding Procedure

The first and second author developed a coding scheme that included the following variables: (a) author(s) and the year of publication, (b) number of participants, (c) \DIC description, (d) participants' age or school grade level, (d) participants' other characteristics such as ability level, SES, El, etc., (e) mathematics content involved, (f) research design, (g) independent and dependent variable(s) and their means (M) and standard deviations (SD), (h) effect size (ES), and (i) narrative research findings. Coding reliability was found by calculating the percentage of agreement between two independent raters on 30% of the studies found in the first stage of searching and screening. The inter-rater reliability was 90.91%.

Results

Types of Existing DICs

To address our first research question regarding the availability of existing interactive tools and their effectiveness that were reported in the review literature of (K-12) mathematics CBAs, based on the functions of the DICs reviewed in this study, we categorized the existing DICs into two major types: (a) DICs that provided language or vocabulary assistance, and (b) DICs that enabled students to construct solutions. Four studies were conducted to investigate the impact of implementing interactive tools that provided language or vocabulary glossaries (Calhoon et al., 2000; Cohen, et al., 2017; Cohen et al., 2020; Roohr & Sireci, 2017). Six studies were found under the second category of effectiveness (Adesina et al., 2014; Applegate, 1993; Kong et al., 2018; Jiang et al., 2021; Ninaus et al., 2017; Threlfall et al., 2007).

Language tools. For the effectiveness category, four studies addressed the use of pop-glossary in mathematics assessment for students who have language or reading problems. Roohr and Sireci (2017) compared the effects of two DICs: A pop-up glossary tools and sticker paraphrasing tool for English language learners (EL s). Findings reported that the pop-up glossary did not provide definitions for all the words or phrases in the test, but only for those that were underlined. For example, when a student clicks on the underlined words or phrases, a window appears with a clarification using definitions, pictures, synonyms, or animations. The sticker paraphrasing tool is intended to provide less difficult paraphrasing after clicking on an icon. More than 2000 students participated in the study. The results show that students used the sticker paraphrasing tool more frequently than the pop-up glossary tool (d = -0.34). Regarding the item-level response time in seconds, results showed that all students took longer on the items that provided either of the two DICs compared with the items that did not provide DICs (d = 0.17). Unfortunately, this study did not report student scores, so we do not know if students scored higher or lower on items with DICs than items without DICs, or if students scored higher on items with one DIC than items with the other DIC.

Two other studies specifically examined the effects of the glossary pop-up tool. Cohen et al. (2017) investigated the impact of using pop-up English glossaries with audio through randomized controlled trials. More than 32,000 third- and-seventh-grade students participated in the study. The items were selected based on experts' judgements of the likelihood that EL s may encounter language difficulties when solving the problems. Math items from a field test were randomly selected to be provided with a pop-up English glossary, whereas the other items were not provided with the pop-up glossary. The results showed that the pop-up glossary slightly inhibited students' performance on mathematics assessments. In the subsequent study about the pop-glossaries, Cohen et al. (2020) scaled up the previous study and over 60000 students from 3rd-grade to 11th-grade participated. Results show that the pop-up glossary accommodation overall was effective for mathematics assessment and did not degrade the measured construct. However, ES could not be obtained for these two studies.

The reading-aloud tool is another DIC that falls into the first category of language assistance tools. Calhoon, et al. (2000) investigated the differences between using a teacher-read and a computer-read tool with and without video. They found that teacher-read tools, computer-read, and computer-read tools with video increased students’ scores on mathematics performance assessment in comparison to the traditional PPA without any reading accommodations. No significant difference was found among these three conditions.

Tools for constructing problem solutions. The second category of DICs identified from the reviewed papers deals with enabling or assisting test takers to construct responses or solutions by interacting with prompts from the computer screen. Adesina et al. (2014) investigated a computerized tool that enabled students to construct responses with an interactive digital device. Two panes were provided to students: the problem pane and the answer pane. The problem pane showed the worded problem and the answer pane showed students’ responses. The problem pane allows the exam taker to drag numeric values using one or both hands. The user could touch and hold two numeric values simultaneously. The touch-and-drag feature activated a menu that contained the main four symbols of the arithmetic operations (i.e., +, -, O, ÷). After selecting one of these symbols a numeric keypad and a text box were displayed. The study investigated the usability and applicability of the tool and compared study students' performance using PPAs, with 39 fifth grade participants. Results reported no significant difference between mathematics scores using the DIC and their corresponding scores in the traditional PPA condition, suggesting access to the DIC did not influence students' scores.

Applegate (1993) investigated the impact of a response-construction tool to test children's analogical reasoning. He compared student performance with a DIC condition and a PPA condition on geometry items and involved 24 kindergarteners. In the DIC condition, a joystick was used to enter data in the response portion. The examinee needed to put the cursor anywhere on the top graphic shapes and press the fire button. The examinee could choose a shape or color. If the correct object was chosen, then it was displayed in an empty box. If the incorrect color was chosen, then all objects were shaded with that color. The study revealed that the DIC condition is significantly more difficult than the PPA condition.

Threlfall et al. (2007), Kong et al., (2008) and Jiang (2021) examined the effects of a pull-and-drag tool. The tool in Threlfall et al., (2007) allowed students to select number cards and put them in boxes to perform an arithmetic operation). This study found that changing from PPA to CBA made little difference in testing validity and legitimacy, and also indicated that some PPA items showed poorer validity than the corresponding equivalent CBA items with DICs, possibly because an equivalent CBA version was less cognitive-demanding and made the problem solving process easier, for example, the pull-and-drag DIC might decrease the likelihood of making mistakes such as using numbers that are not mentioned in the question. Kong et al. (2018) compared the response time differences between computers and tablets on items either with a DIC (e.g., drag-and-drop) or without a DIC (e.g., filling- in-the blank, multiple choice). However, this study did not compare either students' accuracy or response time between items with the "drag-and-drop" DIC and items without a DIC. Jiang et al. (2021) studied students' performance on drag-and-drop items in a large-scale assessment (i.e., National Assessment of Educational Progress, NAEP) in the fourth and eighth graders. The results also revealed that students who answered the item correctly spent shorter time to respond compared to their peers who responded incorrectly.

Effects of tools. Mixed results of the DIC effectiveness were reported in two DIC categories. In all studies that compared differences in students' scores between a DIC condition and a non-DIC condition, only one study (Applegate, 1993) used a between-subject design and provided needed data to calculate an ES, whereas all other studies with a within-subject design did not provide needed data to calculate an ES. Consequently, we were unable to conduct a meta-analysis to estimate an overall effect size for each DIC category. Instead, we summarized the mixed findings, which seem to favor a positive effect of language tools -- two studies (Calhoon et al., 1997; Cohen et al., 2020) reported positive effects, one (Cohen et al., 2017) reported negative effects, and one did not report student accuracy scores. There also seemed to be a negative effect of the response construction category DIC, with two non-significant effects (Adesina et al., 2014; Threlfall et al., 2007), and one negative effects (Applegate, 1993).

Roles of Student Characteristics

To explain the varying effectiveness reported in research studies on the same type of DIC, we further examined whether and how student characteristics (e.g., age, grade level, ability level, EL status) and item features (e.g., mathematics subjects) were related to the effectiveness of DICs. Although we could not calculate weighted effect sizes or perform a meta-regression to detect any significant moderating effects, we identified some patterns, including the following: (a) DICs may inhibit young' children' performance whereas support older children's performance, and the positive effects of DICs increase with age or grade level, and (b) DICs seemed to have different effects in different subgroups, and generally speaking, it appeared to produce greater influences, either positive or negative, on students who were in greater need of help. For example, the language DICs had greater influence on ELs than non-EL students (Cohen et al., 2017; Cohen et al., 2020) and EL students also tended to use the language DICs more frequently (Roohr & Sireci, 2017); and across all 6 studies comparing non-DICs with DICs, students with disabilities benefited from the DIC (Calhoon et al. 2000) and general-education students did not show any changes (Adesina et al., 2014; Threlfall et al., 2007). However, the sample with gifted students was inhibited in a DIC condition (Applegate, 1993). Due to the limited number of studies available, we interpreted these patterns as a hypothesis for future research to verify, rather than as validated conclusions from this synthesis. Well-designed DICs could provide accommodations to address the special needs of students with disabilities and Els, and hence offer promise to make the assessments more assessable and equitable for these special student populations. With poorly-designed DICs, disadvantages can be exacerbated with special student populations by creating extra barriers and demanding greater cognitive load.

Effects of Mathematics Domains

Regarding the influence of mathematics domains and item features, the current findings reported mixed effects in both geometry items and numerical items. It is reasonable to expect that students might obtain greater benefits on geometry items than numerical items because geometry assessment items rely heavily on visual presentations for which DIC features may play a more effective role by enabling students' interactions with diagrams. However, three studies (Applegate, 1993; Cohen et al., 2020; Trelfall et al., 2017) that involved a form of geometry assessment suggested mixed findings. This may be partially attributed to the nature of the items intended to be measured, confounded by other variables such as text or perhaps the less-than-satisfactory geometry DIC effects to the very few available DICs. In the reviewed studies no DICs that might be particularly useful for geometry problem solving, such as drawing supplemental lines, rotations etc., were included, suggesting the need for further research in order to gain richer understanding of the potential benefits of well-designed geometry DICs.

Discussion

This study synthesized prior studies that examined the effects of specific Digitalized-Interactive-components (DICs) in computer-based assessment (CBA). Research questions included: What DICs were available in the literature of CBA in mathematics; what were the main features of these DICs, and how effective were they? How did students' age, gender, and ability level influence performance with the interactive tools? And what mathematical content domains were involved, and whether the DIC effectiveness was influenced by mathematics domains or item features? We systematically searched the literature and reviewed 10 studies that met our inclusionary criteria. The small number of studies identified suggested a need for more research in this important field. While CBA has already been widely implemented in high-stakes, large-scale assessments and DIC is an essential component to ensure the effectiveness of CBA, very little empirical data is available as to the effectiveness of DICs. Although in educational practice, there have been greater use of a variety of DICs that have been developed and used in school practice, results of our systematic search revealed that these DICs that are widely used in school-assessment practice were never validated, suggesting that the research base is lagging behind educational practice. We also discovered that 7 out of the 10 studies were published after 2014, including six studies published in 2017 –2021, suggesting that this is a topic gaining increased attention by the research community.

We categorized DICs that were used as a language tool, and DICs that were used to enable students to construct solutions. Both categories of DICs do not reflect the numerous DIC types in actual practice. For example, NAEP assessments, with the touch-screen tablets mode, have made available DICs such as the read-aloud tool, making digital notes, adjusting screen themes, highlighting texts, dynamic texts, interactive maps, embedded digital calculators, interactive mathematical graphs, and virtual simulations (NCES, n.d). However, most of these DICs are not included in any of the reviewed articles. Also, lacking were some sophisticated DICs that can be found in some commercial online assessment programs such as rotating/flipping/transforming images, digital rulers, annotations. Such a mismatch between the limited types of DICs in research papers and the rich resources of DIC options in practice certainly reflects the gap between the research and the rapid development of technology in the assessment industry. Ideally, educational practice should be guided by education theories and validated by evidence from rigorous educational research; however, in actual school practice it seems that the assessment industry has moved much faster than the tools included by researchers in the studies. Earlier literature (Kim, 1998; Logan, 2015; Smolinsky et al, 2020; Wang et al., 2007) shows that some DICs in practice inhibited, rather than supported students' mathematical problem solving during a CBA. Also, there remains a concern that a CBA with poorly designed DICs could create greater assessment biases against disadvantaged students (Pan, 2016). More research is warranted to ensure the development and validation of science-based assessment tools based on educational theories and cognitive principles.

A limitation of this synthesis is the sparsity of research available - only 10 studies found in peer-reviewed journals. Because access to technical reports produced by some testing companies or organizations (e.g., Educational Testing Service, or American Institute of Research) was not available, a decision was not to include these non-peer-reviewed reports. Moreover, insufficient information was available in the research publications, making us unable to calculate an ES for within-subject design studies. Therefore, the DIC effects and possible influences of participant characteristics and item features were presented as an overall summary of the findings of the related reviewed studies, rather than as a result from a rigorous meta-regression. Therefore, cautions need to be applied to the findings, such as DICs may inhibit young' children' performance whereas support older children's performance, and the positive effects of DICs increase with age or grade level, and DICs seemed to have different effects in different subgroups, and generally speaking, it appeared to produce greater influences, either positive or negative, on students who were in greater need of help.

Conclusions

In the field of mathematics assessment, numerous studies were conducted to investigate the impact of implementing computerized tools, but few studies have examined the effectiveness of implementing DICs. We identified 10 studies that implemented a DIC in a mathematics assessment. The DICs were categorized into two groups: Interactive tools that provided language or vocabulary assistance, and interactive tools that enabled students to construct solutions and ideas. DICs may have differential effects on children of varying ages, and appeared to produce greater influences, either positive or negative, on students who were in greater need of help. Further research studies should be conducted to investigate the effectiveness of implementing DICs in assessing mathematics performance of K-12 students of different ages and with varying special needs. The need for continued research that has the potential to gain greater insight into what knowledge can be gained when struggling students have access to appropriate tools is clear.

References

*Adesina, A., Stone, R., Batmaz, F., & Jones, I. (2014). Touch arithmetic: A process-based computer-aided assessment approach for capture of problem-solving steps in the context of elementary mathematics. *Computers & Education, 78*, 333–343. https://doi.org/10.1016/j.compedu.2014.06.015

* Applegate, B. (1993). Construction of geometric analogy problems by young children in a computer-based test. *Journal of Educational Computing Research, 9*, 61-77. https://doi.org/10.2190/F77B-PPYA-V1EX-LWDY

Aspiranti, K. B., Henze, E. E. C., & Reynolds, J. L. (2020). Comparing paper and tablet modalities of math assessment for multiplication and addition. *School Psychology Review, 49*(4), 453–465. https://doi.org/10.1080/2372966X.2020.1844548

Archbald, & Newmann, F. A. (1988). *Beyond standardized testing : assessing authentic academic achievement in the secondary school.* National Association of Secondary School Principals.

Arslan, B., Jiang, Y., Keehner, M., Gong, T., Katz, I. R., & Yan, F. (2020). The effect of drag-and-drop item features on test-taker performance and response strategies. *Educational Measurement: Issues & Practice, 39*(2), 96–106. https://doi.org/10.1111/emip.12326

Ashton, H. S., Beevers, C. E., Korabinski, A. A., & Youngson, M. A. (2006). Incorporating partial credit in computer-aided assessment of mathematics in secondary education. *British Journal of Educational Technology, 37*(1), 93–119. https://doi.org/10.1111/j.1467-8535.2005.00512.x

Bennett, R. E. (1999). Computer-based testing for examinees with disabilities: On the road to generalized accommodation. *Assessment in higher education: Issues of access, quality, student development, and public policy*, 181-191.

Bennett, S., Maton, K., & Kervin, L. (2008). The digital natives debate: A critical review of the evidence. *British Journal of Educational Technology, 39*(5), 775–786. https://doi.org/10.1111/j.1467-8535.2007.00793.x

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16*(3), 191–205. https://doi.org/10.1207/S15324818AME1603_2

Bergner, Y., & von Davier, A. A. (2019). Process Data in NAEP: Past, Present, and Future. *Journal of Educational and Behavioral Statistics, 44*(6), 706–732. https://doi.org/10.3102/1076998618784700

Bergstrom, B. A. (1992). *Ability Measure Equivalence of Computer Adaptive and Pencil and Paper Tests: A Research Synthesis.* Paper presented at the Annual Meeting of the American Educational Research Association (Sa Francisco, CA, April 20-24, 1992).

Bridgeman, B., & Potenza, M. (1998, April 12-16). *Effects on performance on the computerized SAT I: Reasoning test in mathematics*. The annual meeting of the national council on measurement in education, San Diego.

Buzick (2021, June). Exploring item level use of accessibility supports [Paper Presentation]. Virtual Conference of Educational Accommodations: Future directions.

* Calhoon, M., Fuchs, L. & Hamlett, C. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly, 23*, 271-282. https://doi.org/10.2307/1511349

Cheng, I., & Basu, A. (2009). Interactive graphics for computer adaptive testing. *Computer Graphics Forum, 28*(8), 2033–2045. https://doi.org/10.1111/j.1467-8659.2009.01427.x

Chen, J., & Perie, M. (2018). Comparability within Computer-Based Assessment: Does Screen Size Matter? *Computers in the Schools, 35*(4), 268–283. https://doi.org/10.1080/07380569.2018.1531599

Christensen, L. L., Shyyan, V., Rogers, C., & Kincaid, A. (2014). *Audio support guidelines for accessible assessments: Insights from cognitive lab*s. Minneapolis, MN: University of Minnesota, Enhanced Assessment Grant (#S368A120006), U.S. Department of Education.

* Cohen, D. J., Ballman, A., Rijmen, F., & Cohen, J. (2020). On the reliable identification and effectiveness of computer-based, pop-up glossaries in large-scale assessments. *Applied Measurement in Education, 33*(4), 378–389. https://doi.org/10.1080/08957347.2020.1789137

* Cohen, D., Tracy, R., & Cohen, J. (2017). On the effectiveness of pop-up English language glossary accommodations for EL students in large-scale assessments. *Applied Measurement in Education, 30*(4), 259–272. https://doi.org/10.1080/08957347.2017.1353986

Crawford, L., Higgins, K. N., Huscroft-d'angelo, J.,N., & Hall, L. (2016). Students' use of electronic support tools in mathematics. *Educational Technology, Research and Development, 64*(6), 1163-1182. https://doi.org/10.1007/s11423-016-9452-7

Dindar, M., Kabakçı Yurdakul, I., & Dönmez, F. (2015). Measuring cognitive load in test items: static graphics versus animated graphics. *Journal of Computer Assisted Learning, 31*(2), 148–161. https://doi.org/10.1111/jcal.12086

Flowers, C., Kim, D., Lewis, P., & Davis, V. C. (2011). A comparison of computer-based testing and pencil-and-paper testing for students with a read-aloud accommodation. *Journal of Special Education Technology, 26*(1), 1-12. https://doi.org/10.1177%2F016264341102600102

Gallagher, A., Bennett, R., Cahalan, C., & Rock, D. (2002). Validity and fairness in technology- based assessment: Detecting construct- irrelevant variance in an open-ended, computerized mathematics task. *Educational Assessment, 8*(1), 27–41. https://doi.org/10.1207/S15326977EA0801_02

Gambari, I. A., Ezenwa, V. I., & Anyanwu, R. C. (2014). Comparative effects of two modes of computer-assisted instructional package on solid geometry achievement. *Contemporary Educational Technology, 5*(2), 110–120. https://doi.org/10.30935/cedtech/6119

Gulley, Ann P, Smith, Luke A, Price, Jordan A, Prickett, Logan C & Ragland, Matthew F. (2017). Process-driven math: An auditory method of mathematics instruction and assessment for students who are blind or have low vision. *Journal of Visual Impairment & Blindness, 111*, 465-471. https://doi.org/10.1177/0145482X1711100507

Harlen, W., & Deakin Crick, R. (2003). Testing and Motivation for Learning. *Assessment in Education : Principles, Policy & Practice, 10*(2), 169–207. https://doi.org/10.1080/0969594032000121270

Hassler Hallstedt, & Ghaderi, A. (2018). Tablets instead of paper-based tests for young children? Comparability between paper and tablet versions of the mathematical Heidelberger Rechen Test 1-4. *Educational Assessment, 23*(3), 195–210. https://doi.org/10.1080/10627197.2018.1488587

Hensley, K. K. (2015). *Examining the effects of paper-based and computer-based modes of assessment on mathematics curriculum-based measurement*. ProQuest Dissertations Publishing

Hippisley, J., Douglas, G. & Houghton, S. (2005). A cross-cultural comparison of numeracy skills using a written and an interactive arithmetic test. *Educational Research, 47*, 205-215. https://doi.org/10.1080/00131880500104325

Hoogland, K., Tout, D. (2018). Computer-based assessment of mathematics into the twenty-first century: pressures and tensions. ZDM : *The International Journal on Mathematics Education, 50*(4), 675–686. https://doi.org/10.1007/s11858-018-0944-2

Huff, K. L., & Sireci, S. G. (2001). Validity Issues in Computer-Based Testing. *Educational Measurement, Issues and Practice, 20*(3), 16–25. https://doi.org/10.1111/j.1745-3992.2001.tb00066.x

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

* Jiang, Y., Gong, T., Saldivia, L., Cayton-Hodges, G., & Agard, C. (2021). Using process data to understand problem-solving strategies and processes for drag-and-drop items in a large-scale mathematics assessment. *Large-Scale Assessments in Education, 9*(1), 1–31. https://doi.org/10.1186/s40536-021-00095-4

Kim, J.-P. (1999, October 13-16). *Meta-analysis of equivalence of computerized and P&P tests on ability measures* [Conference Session]. The annual meeting of the Mid-Western Educational Research Association, Chicago, IL.

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22*(1), 22–37. https://doi.org/10.1080/08957340802558326

Koong, C., & Wu, C. (2010). An interactive item sharing website for creating and conducting on-line testing. *Computers and Education, 55*(1), 131–144. https://doi.org/10.1016/j.compedu.2009.12.010

* Kong, X., Davis, L., McBride, Y., & Morrison, K. (2018). Response time differences between computers and tablets. *Applied Measurement in Education, 31*(1), 17–29. https://doi.org/10.1080/08957347.2017.1391261

Lee, G., & Weerakoon, P. (2001). The role of computer-aided assessment in health professional education: a comparison of student performance in computer-based and paper-and-pen multiple-choice tests. *Medical Teacher, 23*(2), 152–157. https://doi.org/10.1080/01421590020031066

Lissitz, R. W., & Jiao, H. (Eds.). (2012). Computers and their impact on state assessments: Recent history and predictions for the future. *IAP*

Logan, T. (2015). The influence of test mode and visuospatial ability on mathematics assessment performance. *Mathematics Education Research Journal, 27*(4), 423-441. https://doi.org/10.1007/s13394-015-0143-1

Masters, J., & Gushta, M. (2018). *Using Technology-Enhanced Items to Measure Fourth Grade Geometry Knowledge*.

McFarlane, A., Williams, J., & Bonnett, M. (2000). Assessment and multimedia authoring - a tool for externalising understanding: Assessment and multimedia authoring. *Journal of Computer Assisted Learning, 16*(3), 201–212. https://doi.org/10.1046/j.1365-2729.2000.00133.x

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ, 339*(7716), e78–336. https://doi.org/10.1136/bmj.b2535

Mavridis, A., & Tsiatsos, T. (2017). Game-based assessment: investigating the impact on test anxiety and exam performance. *Journal of Computer Assisted Learning, 33*(2), 137-150. https://doi.org/10.1111/jcal.12170

National Center for Educational Statistics (NCES). (n.d.). *Digitally based assessment – NAEP*. https://nces.ed.gov/nationsreportcard/dba/

* Ninaus, M., Kiili, K., McMullen, J., & Moeller, K. (2017). Assessing fraction knowledge by a digital game. *Computers in Human Behavior, 70*, 197–206. https://psycnet.apa.org/doi/10.1016/j.chb.2017.01.004

Ninness, H. A. Chris, Ninness, Sharon K, Sherman, Sandra & Schotta, Chuck. (1998). Augmenting computer-interactive self-assessment with and without feedback. *The Psychological Record, 48*(4), 601-616.

Olson, L. (2005). Impact of paper-and-pencil, online testing is compared. *Education Week, 25*(1), 14.

Pantzare. (2012). Students' use of CAS calculators--effects on the trustworthiness and fairness of mathematics assessments. *International Journal of Mathematical Education in Science and Technology, 43*(7), 843–861.

Price, B., Steinle, V., Stacey, K., Gvozdenko, E. (2014). *Using Percentages to Describe and Calculate Change* [Conference Session]. The Annual Meeting of the Mathematics Education Research Group of Australasia (MERGA), Sydney, New South Wales, Australia.

Ridgway, J., & McCusker, S. (2003). Using computers to assess new educational goals. *Assessment in Education: Principles, Policy & Practice, 10*(3), 309–328. https://psycnet.apa.org/doi/10.1080/0969594032000148163

* Roohr, K. C., & Sireci, S. G. (2017). Evaluating computer-based test accommodations for English learners. *Educational Assessment, 22*(1), 35–53.

Russell, M., & Moncaleano, S. (2019). Examining the use and construct fidelity of technology-enhanced items employed by K-12 testing programs. *Educational Assessment, 24*(4), 286–304. https://doi.org/10.1080/10627197.2019.1670055

Schatz, P., Neidzwski, K., Moser, R. S., & Karpf, R. (2010). Relationship Between Subjective Test Feedback Provided by High-School Athletes During Computer-Based Assessment of Baseline Cognitive Functioning and Self-Reported Symptoms. *Archives of Clinical Neuropsychology, 25*(4), 285–292. https://doi.org/10.1093/arclin/acq022

Smoline, D. V. (2008). Some problems of computer-aided testing and "interview-like tests". *Computers & Education, 51*(2), 743–756. https://doi.org/10.1016/j.compedu.2007.07.008

Smolinsky, L., Marx, B. D., Olafsson, G., & Ma, Y. A. (2020). Computer-based and paper-and-pencil tests: A study in calculus for STEM majors. *Journal of Educational Computing Research, 58*(7), 1256–1278. https://doi.org/10.1177%2F0735633120930235

Steedle, J., Pashley, P., Cho, Y., & ACT, I. (2020). Three studies of comparability between paper-based and computer-based testing for the ACT. *ACT Research & Policy*. https://www.act.org/content/dam/act/unsecured/documents/R1847-three-comparability-studies-2020-12.pdf

Stickney, E. M., Sharp, L. B., & Kenyon, A. S. (2012). Technology-Enhanced Assessment of Math Fact Automaticity: Patterns of Performance for Low- and Typically Achieving Students. *Assessment for Effective Intervention, 37*(2), 84–94. https://doi.org/10.1177/1534508411430321

* Threlfall, J., Pool, P., Homer, M., & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics, 66*(3), 335–348. https://doi.org/10.1007/s10649-006-9078-5

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219–238. https://doi.org/10.1177/0013164406288166

Wang, T.-H, Kao, C.-H., & Chen, H.-C. (2021). Factors Associated with the Equivalence of the Scores of Computer-Based Test and Paper-and-Pencil Test: Presentation Type, Item Difficulty and Administration Order. *Sustainability, 13*(17), 9548–. https://doi.org/10.3390/su13179548

Yao, D. (2020). A comparative study of test takers' performance on computer-based test and paper-based test across different CEFR levels. *English Language Teaching, 13*(1), 124–133. http://dx.doi.org/10.5539/elt.v13n1p124