Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 22, August 2022 ISSN 1531-7714

Adapting Paper-Based Tests for Computer Administration: Lessons Learned from 30 Years of Mode Effects Studies in Education

Sarah Lynch[1], *University of British Columbia*

In today's digital age, tests are increasingly being delivered on computers. Many of these computer-based tests (CBTs) have been adapted from paper-based tests (PBTs). However, this change in mode of test administration has the potential to introduce construct-irrelevant variance, affecting the validity of score interpretations. Because of this, when scores from a CBT are to be interpreted in the same way as a PBT, evidence is needed to support the reliability and validity these scores (AERA et al. 2014). Numerous studies have investigated the impact of changing the mode of test delivery from paper to computer, not only in terms of their psychometric properties, but also with regard to possible sources of construct-irrelevant variance. This article summarizes the main lessons learned from mode effects studies in education over the past 30 years and discusses some of the questions remaining.

Keywords: computerized assessment, test administration mode, mode effects, educational tests

Introduction

In today's digital age, computers and other electronic devices are commonplace in many regions of the world. In 2019, 47% of households around the world had a computer in their home, and 57% had access to the Internet via a computer or other electronic device (International Telecommunication Union, 2020). This rate was higher in industrialized countries, where approximately 75% had a home computer, and roughly 84% had Internet access via a computer or other electronic device (International Telecommunication Union, 2020). With computers now playing a pivotal role in our daily lives, many educational tests have transitioned from paper-based to computer-based administration, a transition which has been accelerated by recent historical events.

Over the past two years, test users have needed to quickly adapt to a new reality. In early 2020, the global COVID-19 pandemic led to lockdowns and social distancing measures, forcing many schools and testing organizations that had been administering in-person paper-based tests (PBTs) to quickly transition to administering computer-based tests (CBTs). Even tests that had already transitioned to in-person computerized administration were forced to deliver their CBTs remotely. The *Standards for Educational and Psychological Testing* (AERA et al., 2014) state that a rationale is needed for adapting a test to a new mode of administration. Given the COVID-19 safety concerns, there has been very good reason to adapt paper-and-pencil tests for computer administration.

The benefits of CBTs over PBTs are also a major motivation for the transition. Computerized tests are

[1] The author would like to thank Dr. Bruno Zumbo (University of British Columbia) for his guidance and thoughtful feedback on this research, as well as Dr. Anita Hubley and Dr. Ed Kroc (University of British Columbia) for their supportive discussion.

more efficient than their paper-based counterparts because scoring is automated, enabling faster reporting and feedback; administration is better controlled, improving standardization and test security; and more data can be gathered, permitting more sophisticated psychometric analyses (Way & Robin, 2016; Wise, 2018). CBTs may also enhance the validity of inferences made from test scores and the fairness of a test as they permit the inclusion of novel item types and accessibility options, which can allow examinees to better demonstrate their knowledge, skills, or abilities. However, one drawback of CBTs is that, when needing to compare tests administered on paper and by computer, the change in mode may result in comparability issues with PBTs (Wise, 2018). As a result, the comparability of PBT and CBT scores has been a growing area of research for the past 30 years.

It should be noted that the literature in this area contains a multitude of terms and definitions, resulting in some ambiguity. Some terms refer to the device used to administer the test (e.g., paper-and-pencil based tests, computer-based tests, tablet-based tests), while others refer to the technology through which the test information is accessed (e.g., online tests, Internet-based tests). For the purposes of this review, the term computer-based tests will be used to include tests administered via the following devices: desktop computers, laptops, and tablets.

Studies that examine the comparability of PBTs and CBTs are often referred to as comparability studies or mode effects studies. Regardless of test administration mode, test takers should receive comparable scores, and the interpretations and decisions based on those scores should be the same. Thus, many mode effects studies provide valuable insight into how to effectively adapt traditionally administered paper-and-pencil tests for computer administration.

This literature review will examine the trends and lessons learned from comparability studies on paper-based and computer-based tests in education. In this discussion, the term computer-based test refers to linear or fixed length tests delivered via computer, where all examinees receive either the same test containing the same items (although their order may vary) or alternate forms of a test that have been developed according to the same specifications (Association of Test Publishers, 2002). Computer adaptive tests (CATs), where the computer administers items to an examinee based on their responses to previous items, are not discussed.

A Brief History of Computer-Based Tests

CBTs have evolved considerably over the past 50 years. In the early 1970s, computers were mainly used to administer and score educational and psychological tests, but in the mid-1970s, advances in psychometrics, particularly item response theory, shifted the focus to tailoring test items to individual test takers, and the power of computers was harnessed to deliver these adaptive tests (Moncaleano & Russell, 2018). In the 1980s, the continued advancements in psychometric theory and increasing availability of personal computers led to the expansion of CATs and CBTs into educational testing, which continued to flourish into the 1990s, leading to the rising demand for securely delivered standardized CBTs and the establishment of fully-equipped testing centres around the world (Way & Robin, 2016; Zumbo, 2021). During this time, many large-scale educational tests had transitioned from paper-based to computer-based administration, such as the Graduate Record Exam (GRE), the Test of English as a Foreign Language (TOEFL), and the Graduate Management Admission Test (GMAT) (Moncaleano & Russell, 2018; Way & Robin, 2016). In the early 2000s, CBTs were introduced for K-12 standardized testing in the U.S., but because of the differing availability of information and communications technologies (ICT) across schools and regions, these tests had to be offered in both modes (Way & Robin, 2016). By the mid-2010s, increased investment in ICT for schools led to computer-based testing becoming the norm (Moncaleano & Russell, 2018). At this point, many large-scale tests had moved from paper-based to computerized administration, but not all had. In early 2020, the global COVID-19 pandemic accelerated the need to adapt tests for new modes of delivery, specifically for remote online testing. For large-scale testing companies, PBTs needed to quickly be adapted for computer-based administration, or existing CBTs needed to transition from in-person proctored administration in testing centers to remote proctored administration in examinees' homes or workplaces (Zumbo, 2021).

Implications for Validity and Fairness

The adaptation of PBTs to CBTs has implications for validity and fairness. Validity can be defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). One major threat to validity is construct-irrelevant variance, when test scores are affected by variance from constructs other than the one intended to be measured (Messick, 1995). Changing a test's mode of administration has the potential to introduce numerous sources of construct-irrelevant variance and thus affect the interpretation of test scores. For example, taking a test on a computer requires some degree of computer skills, and these skills, or lack thereof, could potentially be captured in an examinee's test score. The specific type of construct-irrelevant variance introduced by the mode of test administration is often referred to as a mode effect. A mode effect in the broadest sense is "any difference found in test performance that is attributed to the mode of administration" (Way et al., 2015, p. 263). Another concern intertwined with validity is fairness. Fairness means that a test "reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population" (AERA et al., 2014, p. 50). When adapting PBTs to CBTs, it is important to consider the test takers and their contexts. No one should be disadvantaged by the mode of test delivery and measures should be taken to ensure the adapted test is a fair measure of each examinee's knowledge, skill, or ability. For example, to mitigate the unfair effects of computer skills on performance, test administrators can provide examinees with a tutorial or practice test items to familiarize them with the CBT interface prior to the official test administration.

Because changing a test's mode of administration has the potential to introduce various sources of construct-irrelevant variance, the *Standards for Educational and Psychological Testing* (AERA et al., 2014) caution against assuming the interchangeability of scores from a PBT and adapted CBT without evidence; thus, evidence should be gathered to support the reliability of test scores and validity of score interpretations when paper-based measures are adapted for computer-based delivery, or when both modes are administered concomitantly.

Guidelines for Best Practices

To address these comparability concerns, the *Standards for Educational and Psychological Testing* (AERA et al., 2014) and the *International Guidelines on Computer-Based and Internet Delivered Testing* (International Test Commission [ITC], 2005) outline best practices when such adaptations are made. Both publications underscore the need to demonstrate the comparability of the two test modes and minimize sources of construct-irrelevant variance. The ITC guidelines (2005) are specific in their recommendations, stating that a PBT and CBT should be comparable in terms of their reliabilities, means and standard deviations; the two versions should be correlated, and should correlate with similar measures; and a CBT should be designed to minimize sources of construct-irrelevant variance. The AERA et al. (2014) provide more general advice, stating that empirical evidence supporting the validity of interpretations and the reliability of test scores of a CBT adapted from a PBT is warranted, and that potential sources of construct-irrelevant variance should be considered.

What these guidelines make clear is that scores from a CBT adapted from a PBT should not be treated as comparable without evidence. As a result, many comparability studies have examined not only the psychometric properties of PBTs and CBTs, but also the potential sources of construct-irrelevant variance.

Literature Review

The aim of this literature review is to summarize the findings of mode effects studies in educational testing over the past 30 years. This builds on an earlier literature review by Leeson (2006) that examined issues in computer-based testing related to participants and technology. Because the design of mode effects studies has important implications for the generalizability of findings and causal inferences made, this review focuses on peer-reviewed studies that used experimental, quasi-experimental, and in some cases mixed methods designs, in order to investigate potential sources of mode effects.

Based on a review of comparability studies on educational tests over the past 30 years, the potential sources of mode effects can be grouped into the following broad and overlapping categories: test

navigation and layout, item characteristics, cognitive processes, raters' scoring, and examinee characteristics. Each of these factors can affect how examinees interact with an item on paper versus a computer and can potentially result in score differences. However, it is important keep in mind that, given people's increased familiarity and comfort with computers over the past decades, the applicability of findings from 30 or even 10 years ago are worth reconsidering. Further research into whether certain potential sources of mode effects still affect examinees in the same way would be valuable.

Test Navigation and Layout

Differences in the navigation and layout of a paper-based versus a computer-based test have the potential to impact test scores. Although attempts are generally made to make the layout as similar as possible when adapting a PBT for CBT administration, more information can fit on a piece of paper than a computer screen, affecting an examinee's navigation through a test. Differences in test navigation and layout that have been frequently investigated are item review and scrolling.

Item Review. One controllable difference in how examinees navigate through a paper-based or computer-based test is the flexibility to review and change their responses to items. Item review is inherent in PBTs but may or may not be permitted in CBTs. Studies on item review have shown that many examinees do indeed change some of their responses when given the opportunity, and that more often than not, their test scores increase as a result (e.g., Papanastasiou, 2015; Revuelta et al., 2003; Vispoel, 2000). Perhaps more importantly, as Vispoel (1998) highlights, allowing item review can increase the validity of test score interpretations when it reduces sources of construct-irrelevant variance such as test anxiety, typos, or errors in comprehending items because the scores would more accurately represent the examinee's ability and would be less contaminated with error. The trade-off is that permitting item review has been found to increase testing time (e.g., Bodmann & Robinson, 2004; Revuelta et al., 2003; Vispoel, 2000).

Several mode effects studies have investigated the impact of permitting or prohibiting item review and have found mixed results. One study by Luecht et al. (1998) investigated the impact of item review using parallel forms of the Comprehensive Basic Sciences Examination (CBSE), a multiple-choice practice exam for medical students. Comparing scores on a PBT and two CBTs (one permitting and one prohibiting item review), they found no significant differences in scores for PBTs and CBTs administered with and without the option to review items; however, in a follow up survey, 20% of students who had taken the CBT that prohibited item review noted it as feature they disliked. In a similarly designed investigation of an undergraduate psychology test, Bodmann and Robinson (2004) found no significant differences in scores for either CBT condition. More recently, in a mode effects study of a multiple-choice TOEFL reading comprehension test, Toroujeni (2021) compared scores across the same three above mentioned testing conditions. While no significant difference was found between the mean scores on the PBT and either of the CBTs, a significant difference was detected between the two CBTs, with mean scores being significantly higher for the CBT that permitted item review. A similar study by Goldberg and Pedulla (2002) of a GRE practice exam found that examinees taking the PBT outperformed those taking the CBT prohibiting review on all three subtests, while examinees taking the CBT permitting review outperformed them on one of the three subtests.

Based on the existing research, it seems that when adapting a test for computer administration, prohibiting item review may negatively impact examinees' test scores and lead to validity and fairness issues. Therefore, it is advantageous to permit item review in CBTs as it can strengthen the validity of score interpretations by reducing sources of construct-irrelevant variance. It can also improve fairness across modes since item review is inherent in PBTs. What is more, permitting item review can improve examinees' test taking experience since it has been shown to be a desired feature of tests.

Scrolling. Another difference in the navigation of PBTs and CBTs is that examinees can easily scan the entire content of a PBT and flip back and forth through its pages, whereas they must often scroll though the content of a CBT. The need to scroll typically occurs in two parts of a CBT: through a stimulus and its associated items, or through a list of item options.

Scrolling through a Stimulus and Items. Because more information can fit on a written page than on a screen, PBT examinees can typically view a stimulus and items all together, either displayed on a single page or placed side by side on two pages of a test booklet. However, CBT examinees often need to scroll up and down, or back and forth to see the stimulus and items on their computer screen. Scrolling is a concern because it has been proposed that readers may be able to remember and find the fixed location of information on a printed page better than on a computer screen since the relative position of the information on a screen moves as one scrolls (Dillon, 1992). Several comparability studies have speculated that scrolling through a stimulus and items may have resulted in mode effects that disadvantaged CBT examinees (e.g., Choi & Tinkler, 2002; Keng et al., 2008; Poggio et al., 2005).

Scrolling versus Paging. Depending on the computer-based testing interface, scrolling may not be the only option to view a stimulus. In some cases, long stimuli can be separated into sections and placed on separate pages, requiring the examinee to click to move to the next page (i.e., paging). In an effort to identify the best way to present extended texts in CBTs, Higgins et al. (2005) and Pommerich (2004) investigated the effects of scrolling versus paging. Examining three testing conditions (a PBT, a CBT with scrolling, and a CBT with paging) for a fourth grade reading test, Higgins et al. (2005) found no statistically significant difference in mean test scores across the three testing conditions but did note that the mean test score for the PBT was 6% higher than the CBT with scrolling. Comparing the same two CBT conditions in tests of science reasoning and reading, Pommerich (2004) noted a significantly higher mean test score for the science reasoning CBT with paging relative to scrolling, whereas no significant difference was detected between the two reading CBT conditions.

The lesson learned from these studies is that when designing a CBT that involves a lengthy stimulus, such as a reading, permitting paging rather than scrolling may reduce a source of construct irrelevant variance and lead to more accurate test scores. However, more current research on this aspect of computerized testing would be beneficial since examinees today are presumably more accustomed to the navigation requirements of CBTs. Twenty to thirty years ago, scrolling was a source of concern because it was believed that spatial awareness differed when reading on paper versus a screen (Dillon, 1992). However, since people's experience reading on screens has increased over the past 20 years, this may no longer be of concern.

Scrolling through Item Options. Scrolling through item options is a fundamental concern because examinees may not realize the need to scroll and may fail to see all of the options. This can lead to errors in measurement. In their investigation of item properties across PBTs and CBTs, several studies have suggested that scrolling may have contributed to mode effects for certain test items. On a state-wide math and language arts test, Keng et al. (2008) conducted a differential item functioning (DIF) analysis using the test mode (PBT vs CBT) as the grouping variable, and observed that two items exhibiting DIF required scrolling through the item options. One math item that contained diagrams in both the stimulus and item options was entirely visible on a single page of the PBT, but only the first two options were visible on the CBT. Because the correct answer was the second option, Keng et al. (2008) speculated that some students may not have realized the need to scroll to see the final two options, resulting in more CBT examinees selecting the correct response. In the language arts portion of the same test, one item involved selecting the best summary of a text from a list of options. They suggested that the CBT examinees may have been disadvantaged by the need to scroll through the options, whereas the PBT examinees could view them all on one page. More recently, Buerger et al. (2019) examined item formats in a large-scale reading assessment. They found that combo box items (i.e., drop down boxes) where examinees had to scroll down through a list of options tended to be more difficult on computer than paper. Likewise, Gu et al. (2020) found that a multiple-select list item in a Chinese test of critical thinking was significantly more difficult on computer than paper. Upon further investigation, they observed that the item on the CBT required scrolling to see all options, whereas the options appeared all together on the PBT.

There is an important lesson to be gained from these studies regarding item presentation on CBTs. When item options are not all visible to examinees and scrolling is required, this can introduce error into a test score. Items that require scrolling due to drop-down boxes, multi-select lists, or lengthy options seem particularly prone to this formatting problem. Thus, care should be taken when designing CBTs to ensure

that all options are visible to examinees at once. If an item cannot be formatted in such a way, using a different item format, or editing the item would help reduce this source of error.

Item Characteristics

Item Format. Another potential source of mode effects is item format. Here item formats are discussed broadly in terms of selected-response (SR) versus constructed-response (CR). Many SR item formats are highly structured, and answers are restricted to selecting a correct option (or options) from a list, or matching pieces of information. CR item formats are less structured and restricted, requiring examinees to write a response ranging from one or two words (e.g., fill-in-the-blank) to an entire text (e.g., essays).

Selected Response versus Constructed Response. In tests that contain both SR and short CR item formats, some studies have found that CR item formats tend to be more susceptible to mode effects. Russell and Haney (1997) conducted a comparability study with middle school students using a test consisting of National Assessment of Education Progress (NAEP) language arts, science, and math items. The test contained primarily multiple choice (MC) items with some short answer items. No mode effects were detected for the MC items, but students taking the CBT performed significantly better on the science and language arts short answer items. Contrary to Russell and Haney's findings, two mode effects studies by Bennett et al. (2008) and Sandene et al. (2005) of the same state-level math test found that items tended to be more difficult on the CBT, and that CR items were more difficult than MC items, with the mean differences for CR items nearly twice as large than for MC items. They noted that the CR items exhibiting DIF all required considerable editing for computer presentation. Similarly, examining DIF across modes of the Partnership for Assessment of Readiness for College and Careers (PARCC), Liu et al. (2016) found that CR math items tended to function differently across modes, with more items on the grades 3-8 tests favoring CBT examinees, and more items on the high school tests favoring PBT examinees.

The inconsistent results from these studies suggest that SR items may be less prone to mode effects than short CR items. Moreover, CR items seem to sometimes benefit PBT examinees and other times benefit CBT examinees; they may also function differently in different subjects. The reasons for these differences remain unknown, but researchers suggest presentation differences between modes could impact examinees' response processes, thus affecting their performance on the item.

Writing Tests. Whereas SR items are typically objectively scored, CR items tend to be subjectively scored by human raters. Thus, to investigate whether differences in CR items, particularly extended written responses, are due to test delivery mode rather than rater bias toward the presentation mode, steps should be taken to disentangle these two potential sources of error. Several studies have mitigated mode-related rater bias by having the handwritten PBT responses typed verbatim on computer and intermixing them with the CBT responses when presented to raters. One such comparability study by Russell and Haney (1997) involving middle school students found that CBT examinees significantly outperformed PBT examinees, with the former writing nearly twice as much and better organizing their responses into paragraphs. Russell and Plati (2001) continued this line of research on a statewide composition test and found again that students taking the CBT wrote longer essays and received higher scores than students taking the PBT. Linking this performance difference to students' computer usage, they concluded that PBTs consisting of extended constructed response items may "severely underestimate the achievement of students accustomed to writing using a computer" (Russell & Plati, 2001, par 1). More recently, Jin and Yan (2017) conducted a comparability study of the College English Test in China using the same approach to prevent mode-related rater bias. They found that, overall, students performed significantly better on the CBT than the PBT, and that when writing on a computer, students produced texts that were considerably longer, contained longer sentences and fewer errors than when using a pen and paper. They also associated these higher scores with higher levels of computer familiarity.

Even though better performance on CBT writing tasks may seem expected given people's familiarity with writing on computer, other studies have found either no performance differences between modes or mixed results. For instance, Sandene et al. (2005) and Horkay et al. (2006) examined the writing portion of the NAEP. Reducing rater bias by double marking a subset of essays and typing several handwritten

responses, they found no significant differences between CBT and PBT writing tasks in terms of mean test scores or length of texts. A more recent study by Chan et al. (2018) found mixed results for an English for Academic Purposes (EAP) writing test. Deeming the comparison of handwritten and word processed texts appropriate by examining rater severity and reliability, they detected no mode effects in overall test scores; however, they did detect mode effects in one of the scoring rubric domains (lexical resources) that favored handwritten texts. Chan et al. (2018) hypothesized that "some writing sub-constructs are being elicited slightly differently under the two modes" (p. 45). Another study with mixed results by Brunfaut et al. (2018) reported on the writing portion of an English language proficiency exam. Although they mentioned that measures were taken to avoid rater-dependence, little detail was given. Comparing PBT and CBT scores across three levels of English proficiency and two task types, they found that students with lower proficiency performed significantly better on one task when writing on paper; however, for the remaining levels and tasks, there were no significant differences between PBT and CBT scores.

Far fewer studies have found that writers perform better on PBTs than CBTs. However, one study of an adult literacy functional writing test by Chen et al. (2011) found such results. After conducting a rater bias analysis and determining there were no significant scoring differences, they observed that adults who took the PBT significantly outperformed those that took the CBT on all three writing tasks. They also noted that for two of the three tasks there were no significant differences in text length, and even though CBT examinees produced longer texts the remaining task, they did not score higher.

Based on the research in writing studies, evidence suggests that computer skills, specifically word processing skills, are one possible explanation for better writing performance on computers than paper. If an examinee has word processing skills, they are likely better able to demonstrate their writing ability on a CBT than a PBT as they can more easily revise and edit their text on computer. Another related explanation for these mode effects is the congruence between mode of learning and mode of testing. Some researchers recommend that the testing mode should correspond to the learning; in other words, if something is learned on a computer, it should be tested on a computer (Clariana & Wallace, 2002). Therefore, in the one study by Chen et al. (2011) that found better performance on the PBT, one wonders if the fact that the participants were older adults who likely learned to write by hand explains the results. Understanding the learning experiences of examinees is perhaps an overlooked aspect of fair and valid testing. Nowadays CBTs are ubiquitous, and it is often assumed that examinees are accustomed to working on computers; however, there are likely some people who are disadvantaged by this mode of testing, and whose scores do not reflect their true ability as a result. We are not yet at a place in time where all people are accustomed to using a computer. As Horkay et al. (2006) so aptly point out, "conducting a writing assessment in either mode alone may underestimate the performance that would have resulted if students had been tested using the mode in which they wrote best" (p. 1), and this can extend to other subjects that require examinees to construct a response. As a result, test users should consider offering examinees a choice of mode in writing tests. To ensure valid and fair testing practices, it may be appropriate to make accommodations for examinees who are accustomed to writing by hand.

Item Content. Based on the studies described in previous sections, it appears that another potential source of mode effects is the content of an item. This has particularly been observed in tests involving graph comprehension and mathematics. Recent research by Boote et al. (2021) investigated mode effects for a test of graph comprehension for MBA students. The test contained items referring to a Venn diagram, a scatterplot, and a divided bar chart. Although no significant differences in overall test scores were found, at the item level they observed that students scored significantly better on the CBT scatterplot items compared to the PBT. Another study involving graduate students by Gu et al. (2006) aimed to explain DIF in GRE math items by examining item content in terms of "a) verbatim page layout; b) mathematical notation; c) GRE item classifications; and d) mathematical content." (p. 9). Although the overall raw scores did not differ greatly between modes, over 75% of the items were flagged for DIF, with some favoring the PBT mode, and others favoring the CBT mode. Analysis of the item content led Gu et al. (2006) to speculate that items involving arithmetic may be more

difficult on computer, whereas items involving variables and equalities/inequalities may be more difficult on paper. They conjectured that examinees may use different cognitive approaches when responding to items in different modes. Studies of large-scale K-12 math tests have found similar results. In the math portion of the PARCC assessment, Liu et al. (2016) observed that the items most frequently flagged for DIF were those requiring examinees to draw a graph or show their work. For grades 3-8, more items favoured the CBT group, while for high school, more items favoured the PBT group. Keng et al. (2008) found significant mode differences favoring the PBT group in some math items that involved graphing and geometric manipulation. They surmised that items requiring examinees to draw or label graphs may be more difficult on a CBT than a PBT, and that transposing graphs onto scratch paper added an additional step for CBT examinees that may reduce accuracy.

Based on the existing research, it seems that for tests involving math or graphs, some item content may lead to different response processes and performance for examinees on computer versus paper. However, it is not understood why. Based on a meta-analysis of K-12 comparability studies, Kingston (2009) suggested that responding to math items on a computer requires test takers to switch focus between the computer and their scratch paper to answer questions, whereas those taking the test on paper can do so in the question booklet, requiring less change in focus. This may explain some of the cases of superior performance on PBT items, but not the cases of superior performance on CBT items. Further studies on students' responses processes across test delivery modes for different types of mathematical content are needed to explain these differences so that better computerized tests can be created.

Response Processes

Comparability studies can provide valuable insight into examinees' and raters' response processes. Response processes are "the mechanisms that underlie what people do, think, or feel when interacting with, and responding to, the item or task and are responsible for generating observed test score variation" (Hubley & Zumbo, 2017, p. 2). Although research on test navigation and layout, and item characteristics provide insight into response processes, other studies have examined response processes more directly in terms of examinees' cognitive processes and raters' scoring.

Cognitive Processes. While the majority of mode effects studies focus on the comparability of test scores across delivery modes, others have examined the comparability of cognitive process for complex tasks such as math, reading, and writing. These studies shed some light on examinees' response processes and whether they are impacted by testing mode.

Examining cognitive processes in math, Johnson and Green (2006) analysed primary school students' test scores and written work, and also conducted observations and interviews with a subsample of students. They found that about one-third of the students engaged differently with math items on computer versus paper, using slightly different working methods depending on the mode. However, overall test scores were not significantly different between modes. In reading, Kobrin and Young (2003) explored the cognitive processes and test taking strategies of university students for GRE reading passages using think aloud protocols. They found that students engaged in the same test taking strategies and most of the same cognitive processes regardless of mode. Numerous studies in second language writing have analyzed writers' cognitive processes across modes. Weir et al. (2007) examined writers processes via a questionnaire while taking an EAP writing test and found no significant differences in terms of scores or cognitive processes. Using the same questionnaire, Jin and Yan (2017) investigated the writing processes of Chinese students taking the College English Test. Although students were found to engage in similar cognitive processes when writing on paper and computer, they scored significantly higher on the CBT. Research by Li (2006) used think aloud protocols with university students taking an EAP writing test and observed that when examinees wrote on computer, they paid greater attention to higher-order thinking skills and made significantly more revisions to their texts compared to when they wrote by hand. In terms of scoring, the CBT and PBT writing tasks received similar scores across the analytic rubric domains, except for argumentation, in which examinees did better on computer than paper. More recently, Chan et al. (2018) investigated the writing processes of undergraduate students taking an EAP writing test via a questionnaire and interviews. Although the questionnaire did not reveal differences in writing

processes between modes, the interviews highlighted some differences. When writing on paper, some examinees reported more detailed planning, more careful consideration of words and sentence structures, and more revisions at the word level; when writing on computer, students did not feel the need to start with a strict plan, focused more on expressing and organizing ideas at the paragraph and sentence level, did more revising during and after their writing, and made more changes at the sentence and clause level to improve coherence. While overall test scores between modes were not significantly different, scores in one of the rating scale categories (lexical resources) were significantly higher when examinees wrote on paper.

These studies suggest that examinees may engage in similar but slightly different cognitive processes when doing math, reading, or writing on computer versus paper. In most cases, it seems that these variations in processes do not affect performance. However, the number of studies on cognitive processes is limited, and more research is needed to support these findings. What is noteworthy is the difference in information produced by questionnaires versus think aloud protocols. While the questionnaires indicate which processes examinees reported engaging in, the think aloud protocols provide more detailed information about when these processes occur and how often.

Raters' Scoring. The procedures for assigning scores should be the same for PBTs and CBTs. For objective test formats, where the correct response is specified and does not require judgment, the scoring is typically not affected by the mode of test administration. In contrast, for subjective test formats, where a human rater must make a judgment and assign a rating, the scoring is more subjective and may be affected by the test mode. One risk with paper-based versus computer-based subjective test formats, such as writing tests, is that a rater may perceive and score the same text differently depending on whether it is handwritten or typed (Way & Robin, 2016). Several studies have compared the assigned ratings on handwritten versus typed texts, and have found that raters awarded higher scores to handwritten texts (e.g., Breland et al., 2005; Powers et al., 1994; Russell & Tao, 2004), while others have found no significant difference in the mean ratings assigned (e.g., Chan et al., 2018; Coniam, 2006; Johnson et al., 2010).

These conflicting results on raters' scoring suggest that if high-stakes tests are to be administered in dual modes, there is the chance that raters' scoring will be biased. To avoid this potential mode effect, it may be fairer to have the handwritten texts typed prior to being presented to raters.

Examinee Characteristics

Other possible sources of construct-irrelevant variance that have been explored are related to examinee subgroups. In particular, the effects of examinees' computer skills and demographic characteristics have been examined in various mode effects studies.

Computer Skills. Some of the most commonly investigated sources of construct-irrelevant variance are examinees' computer skills. Computer skills have been conceptualized differently in the literature (e.g., computer familiarity, computer use, hands-on skills) and measured differently (e.g., questionnaires, hands-on exercises, or a combination of the two). Nonetheless, it is logical to presume that a person's computer skills could affect their performance on a CBT. If a person has little experience with computers, they may perform less well on a CBT than a PBT. The opposite would likely be true for a person with advanced computer skills. Moreover, it is reasonable to presume that computer skills would affect performance on CR items more than SR items since the former require word processing skills and the latter involve basic interaction with a mouse, keyboard, touchscreen, or touchpad. Thus, in addition to the construct intended to be measured, computer skills may also be reflected in an examinee's test score.

Some studies of tests that contained primarily SR items have found that prior computer use did not contribute to performance differences between PBTs and CBTs (e.g. Higgins et al., 2005, 2010), while other studies of tests that contained both selected and constructed response items have noted that increased computer familiarity was associated with higher CBT scores (e.g., Bennett et al., 2008; Chan et al., 2018; Goldberg & Pedulla, 2002; Horkay et al., 2006; Jin & Yan, 2017; Sandene et al., 2005). It is worth noting that most of these studies provided tutorials to examinees prior to test administration to familiarize them with the CBT interface.

Based on the existing research, it seems that test performance may depend on the degree of computer

skills required to respond to the test items. In tests that are primarily SR or that require a very short CR, the potential effect of computer skills can likely be mitigated by providing a tutorial on the CBT interface and the opportunity to practice. However, for tests that require moderate or extended written responses, the demands are different and require greater word processing skills; thus, for such tests, computer skills can have a greater impact on performance. Therefore, if transitioning from PBT to CBT administration, it is important to consider the demands of the CBT items and examines' computer skills. Only then can appropriate measures be taken to ensure the validity of test score interpretations and fairness to examinees.

Demographics. Another examinee characteristic that is often investigated in mode effects studies is demographics, in particular gender and race/ethnicity. Numerous studies have examined either one or both of these demographic characteristics and found no significant interactions with mode of test delivery (e.g., Bennett et al., 2008; Horkay et al., 2006; Kroehne et al., 2019; Randall et al., 2012; Sandene et al., 2005; Steedle et al., 2020). However, a few studies have noted significant differences in terms of gender (e.g., Boote et al., 2021; Hamhuis et al., 2020) or race/ethnicity (e.g., Chen et al., 2011; Gallagher et al., 2002). However, there do not appear to be obvious patterns related to test subject or item type.

Summary

Table 1 provides a summary of the findings of this literature review of mode effects studies in education over the past 30 years. What is clear from these results is that it is misguided to assume that performance is comparable across modes; thus, when tests are offered in both CBT and PBT modes, or when scores from CBTs adapted from PBTs are to be compared, research should be conducted to support the validity of score interpretations across modes. Moreover, these mode effects studies provide test developers with important considerations and best practices when adapting PBTs for CBT administration.

Discussion

Because differences across testing modes are inevitable, potential sources of construct irrelevant variance should be identified and investigated. As Lenhard et al. (2017) point out "the extent that a measure is invariant across test media depends on the specific measure in question, the participants, and the software and hardware used for testing." (p. 429). The aforementioned studies examined the most commonly identified potential sources of mode effects. However, there are certainly more sources that could be examined. It seems that as more potential sources of mode effects are identified, or the more researchers speculate as to possible causes of mode effects, more investigation is needed. Furthermore, as computers have become more prevalent in our lives, it is worth re-examining findings from older studies to see what results remain consistent, and those that may have changed. As Clariana and Wallace (2002) advise, "As students become as familiar with computer-based testing as they are with paper-based testing, the test mode effect should decrease or disappear" (p. 599). Computers have become increasingly used in our daily lives. Thus, it begs the question: do the findings of studies from 30, 20, 10 or even 5 years ago still hold true today? More up-to-date research on some of these findings would be valuable.

Even after decades of studies on the comparability of PBTs and CBTs, many questions remain. The studies that either suggested or investigated scrolling as a source of error in CBTs are over 15 years old. Now that so many people are accustomed to using electronic devices such as computers, tablets, and cell phones, it seems reasonable to imagine that scrolling may no longer be an issue. Hence, does scrolling through a stimulus and items still disadvantage some CBT examinees, or have most people become accustomed to this type of navigation? Is paging still advantageous over scrolling, or have people become more accustomed to reading on screens? If some people are still disadvantaged by scrolling, who are they?

It would also be useful to understand why some item formats function differently across modes. For short CR items, is it a matter of item presentation, computer skills, rater bias, or changes in response processes? In many studies, these sources of mode effects are not disentangled. For writing tests, it is unclear why, in some instances, writers who produce longer texts and spend more time editing and revising texts composed on computer do not always score significantly higher than when composing on paper. If CBT and PBT scores are similar, is it because the

Table 1. Summary of Findings from Mode Effects Studies in Education

Potential Source of Mode Effects	Findings
Test Navigation and Layout	
Item review	No significant mode effects have been found.
	Permitting item review is encouraged to improve the validity of score interpretations and examinees' test taking experience.
Scrolling through stimulus and items	Scrolling through a stimulus and items may result in mode effects that disadvantage CBT examinees.
Scrolling versus paging	Permitting paging rather than scrolling in a CBT interface may reduce a source of error.
Scrolling through item options	CBT items that require examinees to scroll through options seem particularly prone to mode effects.
Item Characteristics	
Item format	SR items tend to be less susceptible to mode effects than CR items.
Writing tests	Writing test performance differences may be attributable to computer skills and/or congruence of learning and testing.
Item content	For tests involving math or graphs, some item content may lead to different response processes and performance for examinees on computer versus paper.
Response Processes	
Cognitive processes	Examinees may engage in similar but slightly different cognitive processes when doing math, reading, or writing on computer versus paper. In most cases, these variations in processes do not affect performance.
Raters' scoring	Mixed results suggest raters' scoring of handwritten versus typed texts may be biased, so measures should be taken to reduce the bias.
Examinee Characteristics	
Computer skills	Test performance may depend on the degree of computer skills required to respond to the test items.
	Providing a tutorial on the CBT interface can mitigate the effect of computer skills for SR and short CR items.
	Because computer skills seem to have a greater impact on extended CR items, understanding the demands of CR items and examinees' computer skills is necessary.
Demographics	Studies have found mixed results on the interaction between mode of testing and gender and race/ethnicity.

differences in analytic rubric domains cancel them out? Does using a holistic rubric make a difference? Also, why do most examinees write more on computer than paper? Is it related to congruence of learning and testing? If some people write better on paper, why is that?

In terms of item content, more research on math items and graphs would be useful. It is still unclear why

some math, algebra, geometry, graph items function differently across modes. Is it the presentation, the response processes required, the tools available, or something else altogether?

Finally, in terms of examinee characteristics, the impact of computer skills is likely different today than even 5 years ago. Thus, is there a difference in computer skills depending on other factors, such as age, or type of computer use (e.g., gamers versus basic users)? It is important to keep in mind that different regions of the world, and even different regions within wealthy nations, have different access to computers, so it would be useful to understand who may still be disadvantaged by CBTs.

In reviewing the literature on comparability studies, one recurring weakness was the lack of description provided. Many mode effects studies seek to identify sources in the test administration procedure that contribute to differences in PBT and CBT scores. However, there are numerous potential sources that need to be controlled in order to disentangle their contributions to performance differences (Kroehne et al., 2019). As a result, it is important that researchers describe their test administration procedure in enough detail so that readers can decide whether aspects of the testing context were controlled sufficiently to make claims about sources of mode effects. As the AERA et al. (2014) state "the conditions under which the data were collected should be described in enough detail that users can judge the relevance of the statistical findings to local conditions" (p. 26). There are a large number of comparability studies that lack sufficient information for readers to judge the relevance of their findings, particularly with regards to classroom testing. In many cases, information about study design and the tests themselves are lacking or missing entirely, putting into question the appropriateness of the comparison and accuracy of the results. Thus, future comparability studies should provide sufficient description of the study design and the tests being compared in order for readers to judge the relevance of their findings. Better and more controlled research is needed to understand what the potential sources of mode effects are and why they impact test performance.

Technology has evolved considerably over the past 30 years and is even very different today than 10 years ago. This evolution will undoubtedly continue into the future, as will the meaning of a computer or electronic device as we understand it today. The same can be said for the field of testing and assessment; it is changing quickly as technology and world events develop. As such, we need to rethink what is considered an adaptation of a test and when we need to be concerned about it. For instance, will a test adapted from a computer or laptop to a tablet require a comparability study? Would a tablet-based test that incorporates a keyboard and mouse be different than one that incorporates a touch screen and stylus? Perhaps it is more about the similarity in the test users' experience than the type of device when taking tests in different modes. The distinction of what is considered an adaptation will need to be clarified as new computerized devices become available.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Association of Test Publishers. (2002). *Testing guidelines: Guidelines for computer-based testing*. https://www.testpublishers.org/assets/documents/CBTGuidelines.pdf

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, *6*(9), 1–39. https://ejournals.bc.edu/index.php/jtla/article/view/1639

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research*, *31*(1), 51–60. https://doi.org/10.2190/GRQQ-YT0F-7LKB-F033

Boote, S. K., Boote, D. N., & Williamson, S. (2021). Assessing graph comprehension on paper and computer with MBA students: A crossover experimental study. *Cogent Education*, *8*(1), 1–21. https://doi.org/10.1080/2331186X.2021.1960247

Breland, H., Lee, Y. W., & Muraki, E. (2005). Comparability of TOEFL CBT essay prompts: Response-mode analyses. *Educational and Psychological Measurement*, *65*(4), 577–595. https://doi.org/10.1177/0013164404272504

Brunfaut, T., Harding, L., & Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing*, *36*, 3–18. https://doi.org/10.1016/j.asw.2018.02.003

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

Chan, S., Bax, S., & Weir, C. (2018). Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test. *Assessing Writing*, *36*, 32–48. https://doi.org/10.1016/j.asw.2018.03.008

Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing*, *16*(1), 49–71. https://doi.org/10.1016/j.asw.2010.11.001

Choi, S. W., & Tinkler, T. (2002). Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting. In *Paper presented at the annual meeting of the National Council on Measurement in Education* (Issue October). https://www.researchgate.net/publication/274713232

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, *33*(5), 593–602. https://web.b.ebscohost.com/ehost/detail/detail?vid=0&sid=0d4001d9-5db4-48b4-bb4c-3dd4bdd09f7f%40sessionmgr102&bdata=JkF1dGhUeXBlPXNoaWImc2l0ZT1laG9zdC1saXZlJnNjb3BlPXNpdGU%3D#db=eric&AN=EJ657890

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening test. *ReCALL*, *18*(2), 193–211. https://doi.org/10.1017/S0958344006000425

Dillon, G. F. (1992). *A comparison of traditional and computerized test modes and the effect of computerization on achievement test performance - ProQuest*. https://www.proquest.com/docview/304020864/abstract/58D3F78BB92B46C9PQ/1?accountid=14656

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. *Journal of Educational Measurement*, *39*(2), 133–147. https://doi.org/10.1111/j.1745-3984.2002.tb01139.x

Goldberg, A. L., & Pedulla, J. J. (2002). Performance differences according to test mode and computer familiarity on a practice graduate record exam. *Educational and Psychological Measurement*, *62*(6), 1053–1067. https://doi.org/10.1177/0013164402238092

Gu, Lin, Ling, G., Liu, O. L., Yang, Z., Li, G., Kardanova, E., & Loyalka, P. (2020). Examining mode effects for an adapted Chinese critical thinking assessment. *Assessment and Evaluation in Higher Education*, 1–15. https://doi.org/10.1080/02602938.2020.1836121

Gu, Lixiong, Drake, S., & Wolfe, E. W. (2006). Differential item functioning of GRE mathematics items across computerized and paper-and-pencil testing media. *Journal of Technology, Learning, and Assessment*, *5*(4), 1–30. www.jtla.org

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

Higgins, J., Patterson, M. B., Bozman, M., & Katz, M. (2010). Examining the Feasibility and Effect of Transitioning GED Tests to Computer. *Journal of Technology, Learning, and Assessment*, *10*(2).

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning, and Assessment*, *3*(4). www.jtla.org

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, *5*(2), 1–39. www.jtla.org

Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In H. A. M. Zumbo B. D. (Ed.), *Understanding and investigating response processes in validation research* (pp. 1–12). Springer. https://doi.org/10.1007/978-3-319-56129-5

International Telecommunication Union. (2020). Measuring digital development. Facts and figures 2020. In *ITU Publications*. https://www.itu.int/en/mediacentre/Documents/MediaRelations/ITU Facts and Figures 2019 - Embargoed 5 November 1200 CET.pdf

International Test Commission. (2005). *International guidelines on computer-based and internet delivered testing*. www.intestcom.org

Jin, Y., & Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. *Language Assessment Quarterly*, *14*(2), 101–119. https://doi.org/10.1080/15434303.2016.1261293

Johnson, M., & Green, S. (2006). On-Line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and Assessment*, *4*(5), 1–34. www.jtla.org

Johnson, M., Nádas, R., & Bell, J. F. (2010). Marking essays on screen: An investigation into the reliability of marking extended subjective texts. *British Journal of Educational Technology*, *41*(5), 814–826. https://doi.org/10.1111/J.1467-8535.2009.00979.X

Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas Assessment of Knowledge and Skills. *Applied Measurement in Education*, *21*(3), 207–226. https://doi.org/10.1080/08957340802161774

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

Kobrin, J. L., & Young, J. W. (2003). The cognitive equivalence of reading comprehension test items via computerized and paper-and-pencil administration. *Applied Measurement in Education*, *16*(2), 115–140. https://doi.org/10.1207/S15324818AME1602_2

Kolen, M. J. (1999). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *International Journal of Phytoremediation*, *21*(1), 73–96. https://doi.org/10.1207/S15326977EA0602_01

Kroehne, U., Buerger, S., Hahnel, C., & Goldhammer, F. (2019). Construct equivalence of PISA reading comprehension measured with paper-based and computer-based assessments. *Educational Measurement: Issues and Practice*, *38*(3), 97–111. https://doi.org/10.1111/emip.12280

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, *6*(1), 1–24. https://doi.org/10.1207/s15327574ijt0601_1

Lenhard, W., Schroeders, U., & Lenhard, A. (2017). *Equivalence of Screen Versus Print Reading Comprehension Depends on Task Complexity and Proficiency*. *54*, 427–445. https://doi.org/10.1080/0163853X.2017.1319653

Li, J. (2006). The mediation of technology in ESL writing and its implications for writing assessment. *Assessing Writing*, *11*, 5–21. https://doi.org/10.1016/j.asw.2005.09.001

Liu, J., Brown, T., Chen, J., Ali, U., Hou, L., & Costanzo, K. (2016). Mode Comparability Study Based on Spring 2015 Operational Test Data. In *Partnership for Assessment of Readiness for College and Careers*. Partnership for Assessment of Readiness for College and Careers. Available from: New Meridian Corporation. e-mail: info@newmeridiancorp.org; Web site: https://parcc-assessment.org.

Luecht, R. M., Hadidi, A., Swanson, D. B., & Case, S. M. (1998). A Comparative Study of a

Comprehensive Basic Sciences Test Using Paper-and-pencil and Computerized Formats. *Academic Medicine*, *73*(10), s51–s53.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. http://psycnet.apa.org/journals/amp/50/9/741.pdf&uid=1996-10004-001&db=PA

Moncaleano, S., & Russell, M. (2018). A historical analysis of technological advances to educational testing: A drive for efficiency and the interplay with validity. *Journal of Applied Testing Technology*, *19*(1), 1–19. http://jattjournal.net/index.php/atp/article/view/131017

Papanastasiou, E. C. (2015). Psychometric changes on item difficulty due to item review by examinees. *Practical Assessment, Research and Evaluation*, *20*(3), 1–10.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment*, *3*(6), 1–30. http://www.jtla.org

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment*, *2*(6), 1–45. https://ejournals.bc.edu/index.php/jtla/article/view/1666

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). They Think Less of My Handwritten Essay If Others Word Process Theirs? Effects on Essay Scores of Intermingling Handwritten and Word-Processed Essays. *Journal of Educational Measurement*, *31*(3), 220–233. https://doi.org/10.1111/j.1745-3984.1994.tb00444.x

Randall, J., Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the comparability of paper- and computer-based science tests across sex and SES subgroups. *Educational Measurement: Issues and Practice*, *31*(4), 2–12. https://doi.org/10.1111/j.1745-3992.2012.00252.x

Revuelta, J., Ximénez, M. C., & Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. *Educational and Psychological Measurement*, *63*(5), 791–808. https://doi.org/10.1177/0013164403251282

Russell, M, & Plati, T. (2001). *Effects of Computer Versus Paper Administration of a State-Mandated Writing Assessment*. https://www.tcrecord.org/books/exec.asp?ContentID=10709

Russell, Michael, & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, *5*(3), 1–20. http://nis.accel.worc.k12.ma.us

Russell, M, & Tao, W. (2004). Effects of handwriting and computer-print on composition scores: A follow-up to Powers, Fowles, Farnum, & Ramsey. *Practical Assessment, Research and Evaluation*, *9*(1), 1. https://doi.org/https://doi.org/10.7275/9g7k-yr32

Sandene, B., Bennett, R., Braswell, J., & Oranje, A. (2005). Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project, research and development series (NCES 2005–457). In *National Center for Education Statistics*. ED Pubs, P.O. Box 1398, Jessup, MD 20794-1398. Tel: 877-433-7827 (Toll Free). http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457

Steedle, J., Pashley, P., & Cho, Y. (2020). Three studies of comparability between paper-based and computer-based testing for the ACT. In *ACT, Inc.* ACT, Inc.

Toroujeni, S. M. H. (2021). Computerized testing in reading comprehension skill: Investigating score interchangeability, item review, age and gender stereotypes, ICT literacy and computer attitudes. *Education and Information Technologies*, 1–40. https://doi.org/10.1007/S10639-021-10584-2

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement*, *35*(4), 328–347. https://doi.org/10.1111/j.1745-3984.1998.tb00542.x

Vispoel, W. P. (2000). Reviewing and changing answers on computerized fixed-item vocabulary tests. In *Educational and Psychological Measurement* (Vol. 60, Issue 3, pp. 371–384). Sage Publications: Thousand Oaks, CA. https://doi.org/10.1177/00131640021970600

Way, W. D., Davis, L. L., Keng, L., & Strain-Seymour, E. (2015). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Ed.), *Technology and Testing: Improving Educational and Psychological Measurement* (pp. 260–285). Taylor and Francis Inc. https://doi.org/10.4324/9781315871493

Way, W. D., & Robin, F. (2016). The history of computer-based testing. In C. S. Wells & M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future* (pp. 185–207). The Guilford Press. https://psycnet.apa.org/record/2016-24406-011

Weir, C., O'Sullivan, B., Yan, J., & Bax, S. (2007). *6. Does the computer make a difference? The reaction of candidates to a computer-based versus a traditional hand-written form of the IELTS Writing component: effects and impact*.

Wise, S. L. (2018). Computer-based testing. In *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 341–344). SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

Zumbo, B. D. (2021). *A novel multimethod approach to investigate whether tests delivered at a test centre are concordant with those delivered remotely online: An investigation of the concordance of the CAEL*. Paragon Testing Enterprises/UBC Paragon Research Initiative, University of British Columbia.

Citation:
Lynch, S. (2022). Adapting paper-based tests for computer administration: Lessons learned from 30 years of mode effects studies in education. *Practical Assessment, Research, & Evaluation*, 27(22). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/22/

Corresponding Author:
Sarah Lynch
University of British Columbia
Vancouver, British Columbia, Canada

Email: slynch02 [at] student.ubc.ca