Paper and Online Testing Mode Comparability: A Review of Research from 2010–2020

Ann Arthur, PhD, Shalini Kapoor, PhD and Jeffrey Steedle, PhD

Recent advances in technology and rapidly expanding access to electronic devices capable of delivering tests have increased the pace at which online testing (also referred to as computer-based testing) is replacing paper testing (also referred to as paper-based testing). Online testing is perceived as advantageous due to factors such as faster scoring and reporting, greater scheduling flexibility, and relative ease of providing accommodations. During this time of transition, it is necessary for many testing programs to offer both modes of administration. When that situation arises, testing programs are responsible for studying the comparability of scores from paper and online test administrations. When testing in different modes is under consideration, the Standards for Educational and Psychological Testing assert that "the user should have sound rationale and empirical evidence, when possible, for concluding that…the validity of interpretations based on the scores will not be compromised" (AERA et al., 2014, p. 144). Empirical research may indicate the presence of a mode effect, which occurs when scores in one mode are higher than scores in the other mode on the same items for students of the same ability. When mode effects are detected, it may be necessary to apply statistical adjustments via equating methods to support claims that scores from paper and online testing can be interpreted and used interchangeably.

Through 2020, online testing for the ACT® test has been available only to State and District testing clients, which are primarily states and districts that administer the ACT to all 11th graders. In the near future, individual examinees who register to take the ACT on a Saturday "national" testing date will have the option of testing online if a nearby test center offers it and has capacity. In advance of offering online testing on a larger scale, ACT conducted three mode comparability studies to better understand the relationship between paper and online test performance (Steedle, Pashley, & Cho, 2020).

This report provides a review of mode comparability research conducted between 2010 and 2020. This literature review was fueled by a desire to understand results from ACT's mode comparability studies and situate them in the broader research literature. The current review focuses on studies conducted between 2010 and 2020 for two reasons. First, the research literature already includes reviews of earlier mode comparability research (e.g., Texas Education Agency, 2008) and meta-analyses (e.g., Kingston, 2009). Second, technology is increasingly integrated into examinees' lives both in and out of school, so examinees' skills and comfort levels associated with using

electronic devices—including for assessments—might be expected to change over time. Thus, more recent mode comparability studies might show different results than earlier studies. This report summarizes prior reviews of mode comparability research, describes recent studies in several content areas (language arts, mathematics, science, social studies, and others), and concludes with a comparison of earlier and recent research.

Prior Reviews and Meta-Analyses

The Texas Education Agency (2008) published an extensive review of mode comparability studies in K–12 education with results separated into four content areas: mathematics, English language arts, science, and social studies. In each content area, the largest number of studies indicated comparability of scores between paper and online testing. Of the studies that did not, there were many more studies favoring paper testing in mathematics (i.e., paper mathematics scores were higher than online mathematics scores). To a lesser extent, the same was true for English language arts, but more studies favored online testing in science. Jeong (2014) reviewed mode comparability studies from 2000–2014 and also presented results by content area. Again, studies showing evidence of comparability across modes were most common. There were similar numbers of studies favoring paper and online testing in language arts, science, and social studies. However, like the Texas Education Agency (2008) review, there were many more mathematics studies favoring paper testing.

Several researchers have applied meta-analysis to examine mode effects across large numbers of comparability studies. Results are commonly reported as effect sizes (d) in standard deviation units, with negative values indicating higher average scores for paper testing. In an early meta-analysis, paper and online scores were highly correlated for timed power tests but less so for speeded tests (r = .97 vs. .72), and the estimated effect size based on all studies was -0.04 standard deviations, which indicated slightly higher paper scores (Mead & Drasgow, 1993). A later meta-analysis of 51 studies indicated that paper and online testing were similarly difficult on average, but online testing was slightly easier than paper testing in a small number of studies involving high school students (J.-P. Kim, 1999).

Wang and colleagues (2007, 2008) conducted meta-analyses of mode comparability studies for mathematics and reading assessments administered to K–12 students. The 2007 mathematics analysis included 44 studies published from 1989–2005. Of those studies, 13 had statistically significant mode effects, and the overall mean was -0.11 standard deviations ($p < .001$). However, when analyzing a subset of 38 studies with homogeneous effect sizes, there was no statistically significant mode effect (d = -0.06, p = .06). Of the 42 reading studies analyzed in the 2008 study, only 12 had statistically significant mode effects, but the meta-analysis effect size of -0.08 was statistically significant ($p < .001$). Again, results indicated that paper scores were slightly higher on average. Kingston's (2009) meta-analysis covered 81 studies published between 1997 and 2007. The estimated mode effect across all studies was -0.01, but there was a

small advantage for online testing in ELA and social studies (0.11 and 0.15, respectively) and a paper testing advantage in mathematics (-0.06).

Note that small effect sizes observed in meta-analyses might be considered negligible by social science conventions (e.g., Cohen, 1988). However, in the context of educational assessment, the potential impact to individual and aggregate student scores must be considered. For example, it would not be acceptable if students or schools were disadvantaged on accountability tests simply because they participated in online testing (e.g., Herold, 2016).

Recent Studies

Table 1 lists the mode comparability studies reviewed and categorizes them according to the direction of the mode effect indicated by results (paper score > online score, paper score < online score, or comparable). Even if a study included multiple results (e.g., at different grade levels or using different analysis methods), it is shown only once per subject area in Table 1, and its placement reflects whatever the bulk of the evidence indicated. For many authors, the conclusion of “comparability” was equivalent to observing no statistically significant mode effects, and Table 1 follows this convention. In some cases, however, the observed mode effect was relatively large in magnitude, even if it was not significantly different from zero. Low estimation precision associated with small samples sizes may have prevented detecting mode effects in such cases.

Table 1. Results of Mode Comparability Studies Since the Year 2010

	Paper score > Online score	Paper score < Online score	Comparable
Language Arts	Backes & Cowen (2019)	Li et al. (2017), high school	Boo & Vispoel (2012), undergrad
	Chen et al. (2011), adults		Brunfaut et al. (2018), young adults
	Hosseini et al. (2014), undergrad		Holzinger et al. (2011), doctors
	Hosseini & Hashemi Toroujeni (2017), undergrad		Khoshsima et al. (2017), undergrad
	Jeong (2014), grade 6		Khoshsima & Hashemi Toroujeni (2017), undergrad
	Jerrim et al. (2018), 15 year-olds		H. R. Kim et al. (2018), undergrad
	D.-H. Kim & Huynh (2010), grade 9		Zeng et al. (2015), grades 3-12*
	H. J. Kim & Kim (2013), grades 10-11		
	Liu et al. (2016), grades 3-8 and high school		
	Lottridge et al. (2010, 2011), grades 8-9		
	Mangen et al. (2013), grade 10		
	Measured Progress (2018), grades 3 & 6		
Math	Jerrim et al. (2018), 15 year-olds		Boo & Vispoel (2012), undergrad
	Liu et al. (2016), grades 3-8 and high school		Hamhuis et al. (2020), grade 4
	Lottridge et al. (2010, 2011), grades 8-9		Jeong (2014), grade 6
	Minnesota Department of Education (2016), grade 11		Li et al. (2017), high school
	Minnesota Department of Education & Pearson (2012), grades 3-8		Moon (2013), grades 4, 8, & 11
Science	Jeong (2014), grade 6		Cagiltay & Yaman (2013), undergrad
			Li et al. (2017), high school
	Jerrim et al. (2018), 15 year-olds		Chua & Don (2013), undergrad
	Lottridge et al. (2010), grades 10-11*		Hamhuis et al. (2020), grade 4
			Herrmann-Abell et al. (2018), grades 4-12
Social Studies	Jeong (2014), grade 6		Karkee et al. (2010), grade 10
	Lottridge et al. (2010), grades 10-11*		Seo & De Jong (2015), grades 6 & 9

Table 1. Results of Mode Comparability Studies Since the Year 2010—continued

	Paper score > Online score	Paper score < Online score	Comparable
Other Subjects			Bayazit & Aşkar (2012), undergrad
			Boevé et al. (2015), undergrad
			Kalogeropoulos et al. (2013), undergrad
			Nikou & Economides (2013), undergrad

* These studies compared different matching methods. Placement in the table represents the bulk of results.

Language Arts

Twenty studies evaluated mode differences for the language arts, which includes knowledge of English or another language, vocabulary, literacy, reading, and writing. Many studies of elementary and secondary education populations found that paper scores were significantly higher than online scores. D.-H. Kim and Huynh (2010) analyzed data from a 9th grade statewide English assessment. For students without learning disabilities, there was a small but significant mode effect favoring paper testing for total scores (d = -0.05, p < .05) and writing scores (d = -0.06, p < .01). Lottridge, Nicewander, Mitzel (2011) examined results of statewide assessments for students in grades 8 and 9. Online scores were lower than paper scores based on a repeated-subjects design (d = -0.20) and a propensity score matching design (d = -0.25), which resulted in about 10% more paper examinees being classified as proficient in English.

In a study of high school students, reading comprehension scores were significantly higher (d = -0.67; p < .001) for paper testing compared to those who took a scanned copy of the test on computers (H. J. Kim & Kim, 2013). On PARCC English language arts tests, scores from paper testing were higher on average in grade 3 (d = -0.22) and grade 9 (d = -0.30) but not grade 7 (Liu et al., 2016). In another study of PARCC ELA exams administered in grades 3–8, students testing online scored 0.24 standard deviations lower than students testing on paper (p < .001; Backes & Cowan, 2019). On statewide assessments of English language arts, paper scores were higher than online scores for grade 3 (d = -0.21) and grade 6 (d = -0.19; Measured Progress, 2018).

Mode effects favoring paper testing have also been observed in adult populations and in other countries and languages. For example, Chen and colleagues (2011) examined functional writing with a diverse sample of 935 adults aged 16 and older. Results indicated that average paper scores were higher than online scores on all three writing tasks (d = -0.47, -0.55, and -0.27). With Norwegian high school students, there was a mode effect favoring paper testing for reading comprehension scores (d = -0.22, p < .05) after controlling for pretest measures of vocabulary, word reading skill, and reading comprehension (Mangen et al., 2013). Jeong (2014) detected a significant mode effect favoring paper testing on a test of Korean language among 6th grade students (p < .01). In two studies of Iranian undergraduate students, paper scores were significantly

higher than online scores on an English achievement test (Hosseini et al., 2014; Hosseini & Hashemi Toroujeni, 2017). Jerrim and colleagues (2018) examined 2015 Programme for International Student Assessment (PISA) field trial data for 15 year-olds from Germany, Sweden, and Ireland. Across the three countries, students who took the online version had average reading scores between 0.15 and 0.20 standard deviations lower than the students who took the paper version.

Several researchers concluded that scores from paper and online testing are comparable, though the studies tended to involve adult participants. Boo and Vispoel (2012) found no significant differences across modes on subtests measuring undergraduates' vocabulary and ability to interpret literary materials. Holzinger and colleagues (2011) compared paper and online reading comprehension for medical professionals and found no significant differences. In studies of Iranian undergraduates, there were no significant differences between paper and online scores for English vocabulary tests (Khoshsima et al., 2017; Khoshsima & Hashemi Toroujeni, 2017). In a study of English language writing skills for European young adults, there were no mode effects for two out of three tests, but a small, significant mode effect favoring paper testing for the third test (Brunfaut et al., 2018). A small study of writing quality among undergraduate students identified no meaningful differences in essay quality between paper and online testing (H. R. Kim et al., 2018). Only one study reviewed involved students in grades 3–12, and results indicated that scores were comparable for most grades when using propensity score matching with nearest-neighbor or optimal matching algorithms (Zeng et al., 2015), but a mode effect favoring paper testing was observed when using the matching method introduced by Way, Davis, and Fitzpatrick (2006).

A mode effect favoring online testing was observed in only one study reviewed for this report. Li, Yi, and Harris (2017) conducted two large studies of the ACT test with a randomly-equivalent groups design. In both studies, online scores were significantly higher than paper scores for the English test (d = 0.15 and 0.17) and reading test (d = 0.32 and 0.18). Note that the online testing condition offered 5 extra minutes of testing time in the first study, which could have contributed to the larger effect. The second study included two writing prompts. Online writing scores were higher than paper scores for one prompt (d = 0.30, $p < .001$), but there was no significant difference between modes for the second prompt (d = -0.02).

Mathematics

In five mathematics studies, examinees scored higher on paper testing compared to online testing. In the study of 2015 PISA field trial data from Sweden, Ireland, and Germany (Jerrim et al., 2018), average paper mathematics scores were higher than online scores by 0.09 to 0.15 standard deviations. On an item level, the difference in proportion correct was more than 0.10 on 7 out of 67 math items. The Lottridge, Nicewander, and Mitzel (2011) study included an end-of-course algebra test for grades 8 and 9. Paper scores were higher than online scores on average for the within-subjects analysis (d = -0.12) and the propensity score matching analysis (d = -0.10). A PARCC mode comparability study detected higher paper scores on the geometry (d = -0.45) and algebra II (d = -0.20) tests using propensity score matching (Liu et al., 2016). In contrast, online scores on the other three tests (grade 5, grade 7, and

algebra I) were higher than paper scores, with effect sizes ranging from -0.06 to -0.37. Examinees performed better on paper in two studies conducted by the Minnesota Department of Education (Minnesota Department of Education, 2016; Minnesota Department of Education & Pearson, 2012). In the earlier study, examinees testing on paper in grades 3–6 answered an average of at least one more item correct, with effect sizes ranging from -0.12 to -0.13. The effect was smaller for grades 7 and 8 (-0.06 and -0.02, respectively). The second study applied propensity score matching to data from 11th grade students, and the average mode effect was -0.24 standard deviations.

Five studies indicated comparability between scores on paper and online math tests. Boo and Vispoel (2012) administered a quantitative reasoning test to undergraduates and found non-significant differences between scores on the paper and online versions. In another study, there were no significant mode effects for students in grades 4, 8, or 11 on a state accountability test (Moon, 2013). The average effect, which favored online testing in grades 8 and 11, was less than 1 item on a test of more than 50 items. Overall, there was no significant mode effect on a mathematics test administered to Korean 6th graders, but there was among female participants (d = -0.48, $p < .05$; Jeong, 2014). In ACT mode comparability research (Li et al., 2017), there was a non-significant mode effect on ACT math scores in the 2014 study (d = 0.05) and the 2015 study (d = 0.02). A recent study compared performance of 4th grade Dutch students on paper and online (tablet) versions of Trends in International Mathematics and Science Study (TIMSS) tests, and no significant mode effects were detected (Hamhuis et al., 2020).

Table 1 lists no studies in which the mathematics mode effect favored online testing. However, one study was not listed in Table 1 due to results that were difficult to categorize. Jerrim (2015) compared scores of 15 year-olds from 32 countries on the PISA mathematics test administered in two modes in 2012. Averaged across all countries, results of the study might support comparability between paper and online scores. Eleven of the 32 countries had significantly lower scores on online testing, 13 had significantly lower scores on paper testing, and the remainder had non-significant mode effects. Most of the differences were less than 10 PISA score points (or 0.10 standard deviations), but the United States, Columbia, and Brazil were notable exceptions. Online scores in those countries were 17, 20, and 25 points higher than the paper scores, respectively.

Science

In three studies, examinees who took science tests on paper outperformed examinees who tested online. Lottridge, Nicewander, and Mitzel (2010) applied propensity score matching to analyze mode differences on an end-of-course biology test taken by 10th and 11th graders, and the average paper scores were 0.22 standard deviations higher than online scores. In the study of Korean 6th graders (Jeong, 2014), science scores from paper testing were significantly higher than online scores ($p < .05$), but when analyzing gender groups separately, the average difference was statistically significant for females (d = 0.45, $p < .05$) but not for males (d = 0.31). The 2015 PISA study also included a science test administered to 15 year-olds in three countries (Jerrim et al., 2018). Average paper scores were higher than online scores by 0.07, 0.11, and 0.25 standard deviations in Sweden, Ireland, and Germany, respectively, but

only the difference in Germany was statistically significant. The mode effect in Sweden was driven primarily by male students, whose online scores were an average of 0.23 standard deviations lower than paper scores.

The literature review identified four studies in the last decade where performance on online and paper science tests was judged to be comparable. In the earliest study, first-year engineering undergraduates were assigned to take a paper or online chemistry test, and the online scores were not significantly higher than the paper scores (d = 0.15; Cagiltay & Yaman, 2013). When undergraduates were randomly assigned to take a biology test on paper or online, paper scores were not significantly higher than online scores on the pretest (d = 0.17; Chua & Don, 2013). In another study, students in grades 4–12 were randomly assigned to take a test about energy on paper or in one of three online conditions with different online interface features (Herrmann-Abell et al., 2018). Hierarchical linear modelling revealed non-significant main effects for two of the online conditions compared to paper. Although online performance was significantly lower in the third condition, which required students to select responses via radio buttons, the authors concluded that paper and online testing offered equivalent measurement. In the most recent study, Dutch 4th graders took both online (tablet) and paper versions of TIMSS science tests with random counterbalancing for order (Hamhuis et al., 2020). On average, paper science scores were slightly higher than online scores, but not to a statistically significant extent. In an analysis that combined math and science results, male and female students performed similarly on paper testing, but female students scored 0.15 standard deviations higher than males when testing online ($p < .05$).

In science, as in other subject areas, only a recent study involving the ACT test revealed higher average test performance for examinees who tested online. As described above, Li, Yi, and Harris (2017) reported results from two separate studies involving high school students randomly assigned to paper and online versions of the ACT. Online testing performance on the science test was significantly higher on average than paper performance in the first study (d = 0.19, $p < 0.001$), but online testers had five extra minutes to complete the test. In the second study, paper and online testers had the same time constraints, and the average difference in performance was negligible (d = 0.01). For that reason, this study appears in the comparable column of Table 1.

Social Studies

A small number of published studied in the past decade examined possible mode effects on social studies assessments. Jeong (2014) administered paper and online social studies tests, and the average paper score was slightly higher than online, but the difference was not statistically significant, nor was the mode effect statistically significant for males or females when analyzed separately (d = -0.11 and -0.24, respectively). The Lottridge, Nicewander, and Mitzel (2010) study also included end-of-course exams for Civics & Economics (10th grade) and U.S. History (11th grade). A propensity score matching analysis revealed a 0.12 standard deviation difference in performance favoring paper testing on the Civics & Economics exam. For U.S. history, students who took the paper test first scored 0.16 standard deviations higher online, but students who tested online first performed no better or worse on paper.

With data from an 8th grade social studies test, Karkee, Kim, and Fatica (2010) compared matching methods for creating samples useful in studying mode effects. The estimated mode effect using internal matching (i.e., using prior social studies test scores) was 0.08 standard deviations favoring online testing; it was 0.02 when using internal plus external matching (i.e., also using prior test scores from other content areas). The estimated mode effects were greater for male students (0.09 and 0.06) compared to female students (0.04 and -0.04), and Asian and Black students exhibited greater mode effects favoring online testing than White students. The authors concluded that "…the test results did not show statistically discernable mode effects based on…student performance" (p. 14).

Similarly, Seo and De Jong (2015) applied propensity score methods to create matched samples of students who took state social studies assessments on paper or online. When analyzing the 6th grade test and the 9th grade test, the test characteristic curves for paper and online testing were very similar, which indicated that students of a given latent ability level would be expected to earn the same observed score on the test regardless of mode. The estimated mode effects, which both favored online testing, were 0.10 standard deviations for the 6th grade test and 0.08 for the 9th grade test. The authors described these effects as "negligible" (p. 106), and subsequent χ^2 tests indicated that the distributions of performance levels (not proficient, partially proficient, proficient, and advanced) were not significantly different between paper and online testing for either test.

Other Subjects

The literature review includes four additional studies that did not fit into language arts, math, or science, or social studies. In general, the authors of these studies concluded that paper and online scores were comparable. For example, Bayazit and Aşkar (2012) randomly assigned undergraduate students to paper and online versions of an instructional design assessment, and paper scores were not significantly higher than online scores (d = -0.32). In this case, small sample size (23 paper and 17 online) would have made it difficult to detect mean differences. Similarly, Kalogeropoulos and colleagues (2013) randomly assigned undergraduates to paper and online versions of a computer programming assessment. Online multiple-choice scores were higher than paper scores, but the differences between the two modes were not statistically significant (d = 0.27). Online scores were significantly higher on both constructed-response sections of the test, but student testing online had the unfair advantage of being able to compile and test their code.

Nikou and Economides (2013) randomly assigned Greek first-year undergraduate students to take assessments for an informatics course on paper, computer, or mobile device. ANOVA revealed that scores were significantly higher for online testing on mobile devices compared to paper testing (d = 0.51, p < .01), but other differences between conditions were not statistically significant. The same pattern in results was observed for females (d = 0.55, p < 0.05), but no mode effects were statistically significant for males. In another study involving first-year undergraduate students, Boevé and her colleagues (2015) randomly assigned students to take either a midterm or final biopsychology exam online. For both exams, the average difference between online and paper scores was non-significant, though the average online score was slightly higher on the final exam (d = 0.09).

Discussion and Conclusions

In earlier literature reviews (Jeong, 2014; Texas Education Agency, 2008), the greatest number of studies supported comparability between scores from paper and online testing, but there were similar numbers of studies showing mode effects favoring paper and online testing. This finding was corroborated by meta-analyses with near zero estimates of mode effects when combining results from numerous studies (e.g., Kingston, 2009). The only notable deviation from this trend occurred in mathematics, where very few studies had results indicating that online testing was easier than paper.

If examinees' experience using computers and taking assessments online has increased over time, one might expect fewer recent studies to show mode effects favoring paper testing. Yet, as indicated by Table 1, the proportion of mode comparability studies in language arts favoring paper testing increased relative to prior literature reviews. In mathematics, it remained very rare for a study to exhibit mode effects favoring online testing. Nearly all mode comparability studies published in the last decade indicated comparability between paper and online scores or that paper was easier than online. Some studies showed non-significant mode effects favoring online testing (e.g., Cagiltay & Yaman, 2013; Moon, 2013), but most of the studies supporting comparability had non-significant mode effects favoring paper testing. Overall, this literature review supports the conclusion that paper testing is often easier than online testing, but not always. Consequently, large-scale testing programs in transition from paper to online testing must evaluate mode comparability and possibly adjust for mode effects to ensure examinees are not disadvantaged by testing mode, especially when there are stakes attached to test performance..

Ultimately, this literature review highlighted the fact that ACT mode comparability studies are outliers in recent mode comparability research. Yet, results from the Li, Yi, and Harris (2017) study were corroborated by ACT mode comparability studies conducted in October 2019, December 2019, and February 2020 (Steedle, Pashley, & Cho, 2020). In those studies, results consistently indicated that examinees who tested online had slight score advantages over those who tested via paper, especially on the English and reading tests. This result is possibly connected to the fact that, unlike tests administered in other mode comparability studies, the ACT test is somewhat speeded, and speededness is known to moderate mode effects (Mead & Drasgow, 1993). For example, examinees who take the ACT online can activate an on-screen timer to display time remaining, and this could help examinees pace themselves more effectively. Considering the current understanding of mode comparability for the ACT test, ACT forms administered in different testing modes will continue to be equated to ensure that ACT scores are comparable regardless of testing mode. ACT will continue to monitor mode effects as more examinees gain access to online testing and update mode adjustment procedures as needed.

References

AERA, APA, & NCME. (2014). *The standards for educational and psychological testing*. Washington, DC: American Educational Research Association. https://www.apa.org/science/programs/testing/standards

Backes, B., & Cowan, J. (2019). Is the pen mightier than the keyboard? The effect of online testing on measured student achievement. *Economics of Education Review, 68*(1), 89–103. https://doi.org/10.1016/j.econedurev.2018.12.007

Bayazit, A., & Aşkar, P. (2012). Performance and duration differences between online and paper–pencil tests. *Asia Pacific Education Review, 13*(2), 219–226. https://doi.org/10.1007/s12564-011-9190-9

Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y., & Bosker, R. J. (2015). Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment. *PloS One, 10*(12), e0143616. https://doi.org/10.1371/journal.pone.0143616

Boo, J., & Vispoel, W. (2012). Computer versus paper-and-pencil assessment of educational development: A comparison of psychometric features and examinee preferences. *Psychological Reports, 111*(2), 443–460. https://doi.org/10.2466/10.03.11.PR0.111.5.443-460

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*(1), 3–18. https://doi.org/10.1016/j.asw.2018.02.003

Cagiltay, N., & Yaman, S.-O. (2013). How can we get benefits of computer-based testing in engineering education? *Computer Applications in Engineering Education, 21*(2), 287–293. https://doi.org/10.1002/cae.20470

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

Chua, Y. P., & Don, Z. M. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. *Computers in Human Behavior, 29*(5), 1889–1895. https://doi.org/10.1016/j.chb.2013.03.008

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Erlbaum.

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,* 1–19. https://doi.org/10.1111/bjet.12914

Herold, B. (2016, February 10). PARCC scores lower for students who took exams on computers. *Education Week*. https://www.edweek.org/ew/articles/2016/02/03/parcc-scores-lower-on-computer.html

Herrmann-Abell, C. F., Hardcastle, J., & DeBoer, G. E. (2018, March). *Comparability of computer-based and paper-based science assessments*. Paper presented at the 2018 annual meeting of National Association for Research in Science Education (NARST), Atlanta, GA.

Holzinger, A., Baernthaler, M., Pammer, W., Katz, H., Bjelic-Radisic, V., & Ziefle, M. (2011). Investigating paper vs. screen in real-life hospital workflows: Performance

contradicts perceived superiority of paper in the user experience. *International Journal of Human-Computer Studies, 69*(9), 563–570. https://doi.org/10.1016/j.ijhcs.2011.05.002

Hosseini, M., Abidin, M. J. Z., & Baghdarnia, M. (2014). Comparability of test results of computer based tests (CBT) and paper and pencil tests (PPT) among English language learners in Iran. *Procedia - Social and Behavioral Sciences, 98*(1), 659–667. https://doi.org/10.1016/j.sbspro.2014.03.465

Hosseini, M., & Hashemi Toroujeni, S. M. (2017). Replacing paper-based testing with an alternative for the assessment of Iranian undergraduate students: Administration mode effect on testing performance. *International Journal of Language and Linguistics, 5*(3), 78–87. https://doi.org/10.11648/j.ijll.20170503.13

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

Jerrim, J. (2015). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice, 23*(4), 495–518. https://doi.org/10.1080/0969594X.2016.1147420

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education, 44*(4), 476–493. https://doi.org/10.1080/03054985.2018.1430025

Kalogeropoulos, N., Tzigounakis, I., Pavlatou, E. A., & Boudouvis, A. G. (2013). Computer-based assessment of student performance in programing courses. *Computer Applications in Engineering Education, 21*(4), 671–683. https://doi.org/10.1002/cae.20512

Karkee, T., Kim, D.-I., & Fatica, K. (2010, April). *Comparability study of online and paper and pencil tests using modified internally and externally matched criteria*. Paper presented at the 2010 annual meeting of the American Educational Research Association, Denver, CO. https://www.measurementinc.com/sites/default/files/2017-05/Online%20and%20Paper%20and%20Pencil%20Comparability%20Study%20with%20Alternate%20Design.pdf

Khoshsima, H., & Hashemi Toroujeni, S. M. (2017). Comparability of computer-based testing and paper-based testing: Testing mode effect, testing mode order, computer attitudes and testing mode preference. *International Journal of Computer, 24*(1), 80–99.

Khoshsima, H., Hosseini, M., & Hashemi Toroujeni, S. M. (2017). Cross-mode comparability of computer-based testing (CBT) versus paper-pencil based testing (PPT): An investigation of testing administration mode among Iranian intermediate EFL learners. *English Language Teaching, 10*(2), 23–32. https://doi.org/10.5539/elt.v10n2p23

Kim, D.-H., & Huynh, H. (2010). Equivalence of paper-and-pencil and online administration modes of the statewide English test for students with and without disabilities. *Educational Assessment, 15*(2), 107–121. https://doi.org/10.1080/10627197.2010.491066

Kim, H. J., & Kim, J. (2013). Reading from an LCD monitor versus paper: Teenagers' reading performance. *International Journal of Research Studies in Educational Technology, 2*(1), 15–24. https://doi.org/10.5861/ijrset.2012.170

Kim, H. R., Bowles, M., Yan, X., & Chung, S. J. (2018). Examining the comparability between paper- and computer-based versions of an integrated writing placement test. *Assessing Writing, 36*(1), 49–62. https://doi.org/10.1016/j.asw.2018.03.006

Kim, J.-P. (1999, October). *Meta-analysis of equivalence of computerized and P&P tests on ability measures*. Paper presented at the 1999 annual meeting of the Mid-South Educational Research Association, Chicago, IL. https://www.learntechlib.org/p/91064/

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

Li, D., Yi, Q., & Harris, D. (2017). *Evidence for paper and online ACT® comparability: Spring 2014 and 2015 mode comparability studies.* Iowa City, IA: ACT. https://www.act.org/content/dam/act/unsecured/documents/Working-Paper-2016-02-Evidence-for-Paper-and-Online-ACT-Comparability.pdf

Liu, J., Brown, T., Chen, J., Ali, U., Hou, L., & Costanzo, K. (2016). *Mode comparability study based on spring 2015 operational test data*. Washington, DC: Partnership for Assessment of Readiness for College and Careers.

Lottridge, S. M., Nicewander, W. A., & Mitzel, H. C. (2010). Summary of the online comparability studies for North Carolina's End-of-Course Assessment Program. In P. C. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (p. 13-32). Washington, DC: Council of Chief State School Officers.

Lottridge, S. M., Nicewander, W. A., & Mitzel, H. C. (2011). A comparison of paper and online tests using a within-subjects design and propensity score matching study. *Multivariate Behavioral Research, 46*(3), 544–566. https://doi.org/10.1080/00273171.2011.569408

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research, 58*, 61–68. https://doi.org/10.1016/j.ijer.2012.12.002

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Measured Progress. (2018). *Massachusetts Comprehensive Assessment System 2017-2018: Mode linking report.* Alpharetta, GA: Measured Progress.

Minnesota Department of Education. (2016). *Equating student scores across test administration modes: Grade 11 mathematics MCA - III* (p. 8). Roseville, MN: Minnesota Department of Education. https://education.mn.gov/MDE/dse/test/Tech/

Minnesota Department of Education, & Pearson. (2012). *Mathematics Minnesota Comprehensive Assessment-Series III (MCA-III): Mode comparability study report.* Roseville, MN: Minnesota Department of Education. https://education.mn.gov/MDE/dse/test/Tech/

Moon, J. L. (2013). *Comparability of online and paper/pencil mathematics performance measures* [Doctoral dissertation, University of Nebraska - Lincoln]. http://digitalcommons.unl.edu/cehsdiss/168

Nikou, S. A., & Economides, A. A. (2013). *Student achievement in paper, computer/web and mobile based assessment*. Paper presented at the Balkan Conference on Informatics, Thessaloniki, Greece. https://www.researchgate.net/publication/265397235_Student_achievement_in_paper_computerweb_and_mobile_based_assessment

Seo, D. G., & De Jong, G. (2015). Comparability of online- and paper-based tests in a statewide assessment program: Using propensity score matching. *Journal of Educational Computing Research, 52*(1), 88–113. https://doi.org/10.1177/0735633114568856

Steedle, J., Pashley, P., & Cho, Y. (2020). *Three studies of comparability between paper-based and computer-based testing for the ACT.* Iowa City, IA: ACT. https://www.act.org/content/dam/act/unsecured/documents/R1842-paper-online-testing-modes-2020-12.pdf

Texas Education Agency. (2008). *A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based tests*. Austin, TX: Texas Education Agency. https://tea.texas.gov/sites/default/files/2008-LiteratureReviewComparabilityReport.pdf

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in brade K-12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219–138. https://doi.org/10.1177/0013164406288166

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5–24. https://doi.org/10.1177/0013164407305592

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). *Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills*. Paper presented at the 2006 annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Zeng, J., Yin, P., & Shedden, K. A. (2015). Does matching quality matter in mode comparison studies? *Educational and Psychological Measurement, 75*(6), 1045–1062. https://doi.org/10.1177/0013164414565006

Ann Arthur, PhD
Ann Arthur is a psychometrician in the Assessment Transformation department. Her research interests include mode comparability, process data, and latent variable models.

Shalini Kapoor, PhD
Shalini Kapoor is a senior psychometrician in the Assessment Transformation department. Her research interests include score comparability, equating, automated test assembly, and computerized adaptive testing.

Jeffrey Steedle, PhD
Jeffrey Steedle is a lead psychometrician in Assessment Transformation directing the team responsible for statistical analyses for the ACT test and guiding research studies related to maintaining measurement quality while making changes to the assessment program. Jeff holds advanced degrees in education, statistics, and educational psychology, and his research interests include assessment validation and motivation on achievement tests.