# A Comparison of Paper-based and Computer-based Formats for Assessing Student Achievement

***Meghan B. Scrimgeour***
***Haigen H. Huang***
**Wake County Public School System**

*Given the growing trend toward using technology to assess student learning, this investigation examined test mode comparability of student achievement scores obtained from paper-pencil and computerized assessments of statewide End-of-Course and End-of-Grade examinations in the subject areas of high school biology and eighth-grade English Language Arts and math. Propensity score matching was used to generate comparable groups of students who were assessed using paper-pencil or computer-based formats. T-tests and generalized linear models were further used to examine test mode effect. Analyses revealed a small test mode effect for all three subjects such that students using the paper-based format achieved higher scores than students using the computer-based format. The findings are germane to school districts transitioning to computerized assessments and investigating test mode comparability*.

Over the past three decades, there has been a growing trend in public education towards transitioning from traditional paper-based assessments to computer-based assessments of student achievement. In fact, 2015-16 was the first academic year during which the majority of U.S. state-required summative assessments in Grades 3-8 were delivered via a technology format (online/computer-based) in contrast to traditional paper-and-pencil format (EdTech Strategies, 2015). As of 2016, roughly two dozen states administer K-12 state assessments online (Backes & Cowan, 2019; Farmer, 2016), and the two consortia of Common Core-based tests, Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced, are transitioning to computer-based testing (Backes & Cowan, 2019). This trend is occurring at both the elementary and secondary education levels as well as in all main content areas (i.e., reading, math, science, and social studies; Bennett, 2003). North Carolina is among the states administering significant numbers of tests online and is substantially increasing that number each year. During the 2016-17 school year, 1.7 million assessments were completed online (NCDPI, email communication, May 13, 2019). During the 2017-18 school year, the number of online assessments increased to 2.1 million, and during the fall 2018, 90.5% of all assessments were completed online (NCDPI, email communication, May 13, 2019).

The transition to online testing has been driven by several advantages that affect both the states conducting the assessments and the students completing them. These strengths include increased flexibility in designing test items, access to a large repository of items, efficient administration, immediate scoring and reporting of results, reduced measurement errors, reduced testing costs, increased student motivation and engagement, improved accessibility for students with special needs, reduced opportunities for student and staff cheating, and consistency across classroom activities and assessments in supporting students' computer literacy skills which are essential for both attaining and maintaining jobs (Backes & Cowan, 2019; Bennett, 2003; Boo & Vispoel,

2012; EdTech Strategies, 2015; Ghaderi et al., 2014; Kim & Huynh, 2010; Randall et al., 2012; Thurlow et al., 2010; U.S. DoE, 2013).

These advantages, however, are tempered by several challenges. Schools' insufficient capacity and infrastructure (e.g., technology devices and connectivity) to administer assessments to all students at once serves as one of the primary drawbacks (Randall et al., 2012; Thurlow et al., 2010; U.S. DoE, 2013). To accommodate for the lack of devices, testing windows are extended thereby increasing the opportunity for students and staff to widely and rapidly disseminate test-item knowledge to students who have not yet taken the test (U.S. DoE, 2013). Challenges also include costs associated with equipping schools with the needed technology, a lack of school staff available to keep equipment running, technical difficulties attributed to the assessment providers and their software, required staff training in order to administer tests with fidelity, and security threats (Davis, 2014; Thurlow et al., 2010; U.S. DoE, 2013).

Students' performance on standardized assessments has important implications not only for the students themselves, but also their teachers, schools, districts, and communities. For example, achievement scores inform identification of students for gifted and talented programs, consideration for special education programs, grade promotion and retention, course placement, student graduation, improvements in instruction, targeted interventions, teacher evaluations, school accountability determinations, distribution of school resources, families making residential location decisions, and researchers' regular use of student test scores as an outcome measure (Backes & Cowan, 2019; Duque, 2016; U.S. DoE, 2013). During the transition from paper-and-pencil assessments to computerized assessments, a period of time is expected during which both methods of administration are used concurrently (Boo & Vispoel, 2012; Randall et al., 2012; Kim & Huynh, 2010). If students' test scores are compared over time and/or if scores are aggregated across students when some students have completed the assessment on paper and others on computer, it is paramount that test mode comparability has been established (APA, 1986; Bennett et al., 2008; International Test Commission, 2005). If students' achievement scores are not equivalent across assessment delivery modes, the ability to draw valid conclusions may be reduced and school-based decisions may be incorrectly informed.

Moreover, standards for educational and psychological testing (AERA, APA, NCME, 2014, Standards 9.7 & 9.9; APA, 1986) stipulate that comparability studies need to be conducted if tests are administered in different modes and that interpretation of students' assessment outcomes shouldn't be influenced by the mode of test administration or the device used to access test content (Davis et al., 2017; Lottridge et al., 2011; OAERS, in progress). Variations in how test information is presented to students as well as how students interact with that information must be taken into consideration when interpreting and using assessment results (DePascale et al., 2016). Therefore, when transitioning from paper-based assessments to computerized assessments, districts should examine test mode comparability of scores (i.e., students who take a test on a computer should receive the same score if they take the same test on paper) to ensure that bias is not introduced into students' performance due to factors independent of their academic knowledge and skill levels.

Although districts are increasingly moving towards computerized assessments, research findings examining test mode comparability across paper- and computer-based testing formats have been

inconsistent. While some studies have shown test mode comparability (e.g., Lottridge et al., 2011; OAIS, 2007; Poggio et al., 2005; Wang, 2004) others have shown small differences in test mode administration (e.g., Backes & Cowan, 2019; Bennett et al., 2008; Pomplun et al., 2006; Russell, 1999; White et al., 2012). For example, across two investigations, middle and high school students took End-of-Course (EOC) biology, algebra, and English tests in both computer- and paper-based formats. Results showed that, for biology, there was not a significant difference in mean scores across the two modes of administration. For algebra and English, however, students scored higher on the paper-based tests than on the computer-based tests though these effects were small (Kim & Huynh, 2007; 2008). Comparing eighth-grade students' performance on a writing test administered either on paper or computer, Horkay et al. (2006) found no significant mean score differences between the two assessment modes. In contrast, in their investigation of eighth-grade students' performance on either a computer- or paper-based math test, Bennett et al. (2008) found that students scored higher on the paper-based math items in comparison to the computer-based items. It is important to note that these mode differences in general were small. Finally, in an examination of third through eighth-grade students' performance on math and English Language Arts (ELA) state-mandated exams, results indicated test mode differences where students across all grades scored higher on both the math and ELA paper-based exams in comparison to students who took the computer-based exams (Duque, 2016).

Several meta-analyses have also concluded that in general, any mode effects found between paper and computerized test administrations tend to be either non-significant or small in effect size. For example, Wang et al.'s (2007; 2008) meta-analyses of K-12 computer and paper math and reading tests found no significant difference in mean performance between the two testing modes. Furthermore, Paek's (2005) review of comparability studies suggested test mode equivalency across grades and academic subjects, and when test mode differences were detected, they were small in effect size or statistically insignificant. In comparison, Kingston's (2009) research synthesis of test scores in Grades 1-12 showed that computer-based administration provided a small advantage for ELA and social studies tests, but paper-based administration provided a small advantage for math tests. Moreover, in their summary of studies conducted with K-12 state departments of education, Way et al. (2008) concluded that students' performance on paper and online science tests was likely comparable. However, comparability results for math, reading, and social studies were less clear and more complicated to interpret. When mode effects were found across subjects, they were small. Although the burgeoning literature is shedding light on increased computerized testing in schools and the accompanying question of test mode comparability, this body of work overall suggests mixed results.

Students' differing achievement scores across modes of testing administration may be attributed to several factors such as: presentation characteristics (e.g., number of items that fit on a computer screen vs. a printed page of a test booklet, font type and size, line spacing, computer screen resolution), response requirements (i.e., paper-based exams require knowing how to pencil-in multiple-choice response bubbles and hand writing responses to open-ended questions whereas computer-based exams require knowing how to use technology to point, click, scroll, type, drag, drop, select drop down menu items, etc.), and general administration characteristics (e.g., adaptive vs. fixed form such that computer-based exams may be adaptive, and timed vs.

untimed such that paper-based exams typically require students to wait until the allotted time has elapsed before proceeding to the next section; Bennett, 2003; Duque, 2016; Leeson, 2006).

Another factor that may differentially affect test mode comparability is students' accessibility to and experience using technology. Computer-based tests may measure students' proficiency in computer literacy, and if students have varying levels of familiarity with technology, their experience interacting with the exam and recording responses may differ (Backes & Cowan, 2019). For example, Bennett et al. (2008) found that eighth-grade students' performance on a computer facility test predicted their performance on the computer-based math test after controlling for math proficiency thereby suggesting that students' familiarity with computers may impact their performance when taking computer-based math tests. Similarly, computer familiarity/proficiency predicted computer-based writing test performance with the result that eighth-grade students with more hands-on computer skills scored higher on a writing test than students with less skills (Horkay et al., 2006). Choi and Tinkler (2002) assessed third and 10$^{th}$-grade students with multiple-choice reading and math tests delivered on paper and by computer. Results showed that computer-based reading and math tests were more difficult for third-grade students, but the paper-based version was more difficult for 10$^{th}$-grade students. This suggests that taking computer-based assessments may be more of a novelty to younger students in comparison to older students. More exposure to and experience with computers may aid in reducing test mode effect.

Extant research also suggests that during the first year of online testing, a temporary adjustment to the new testing format and students' unfamiliarity with navigating technology devices to complete testing may account for a portion of mode effects. For example, Backes and Cowan (2019) examined PARCC math and ELA achievement data for students in third through fifth grades. Overall, results revealed test mode effects such that students who took the computer-based exams scored about 0.10 standard deviations lower in math and about 0.25 standard deviations lower in ELA than students taking the paper-based exams. Between year one and year two of testing, results began to show evidence of fadeout of test mode effects. Test mode effects for second-time test takers were about one third as large as the first year in math and about half as large in ELA. Collectively, this compilation of results may therefore lend themselves to inform testing policy so that districts may want to exercise caution when interpreting and using transition-year scores for accountability purposes (Backes & Cowan, 2019). If students' scores are lower due to testing delivery format as opposed to a lack of knowledge or skill, this shouldn't directly impact important decision for students, teacher evaluations, or school accountability. It is suggested that over time, schools and districts will improve their ability to administer computer-based assessments and students will become more familiar with the user interface thereby correcting the possible issue of test mode effects.

Given the mixed research findings on test mode comparability, it is challenging to make use of that research to inform local district decision-making in implementing computer-based assessments. Therefore, in the current study, we used a quasi-experimental method—propensity score matching—to examine recent data from a large urban school district in order to address test mode comparability for several subjects and across multiple grades and academic years. The following research questions were addressed: Is there test mode comparability between paper-pencil and computer-based statewide standardized assessments? If a test mode effect exists, does

it vary across academic subjects? By addressing these questions, this study aims to contribute to the existing research and thereby inform and encourage school districts' exploration of test mode comparability.

## Method

### Data and Sample

To examine test mode comparability, the current study used data drawn from a large urban school district in the southeastern United States. High school and middle school students' performance on state-mandated standardized End-of-Course (EOC) and End-of-Grade (EOG) assessments were utilized. The EOC data included students' achievement scores on the high school biology test for the 2015-16, 2016-17, and 2017-18 academic years. The EOG data included students' achievement scores on the eighth-grade ELA and math tests for the 2016-17 and 2017-18 academic years.

Each school in the district decided whether their students would complete the assessments using either the paper-pencil or computerized delivery format. As a result, this sample consisted of two non-randomly assigned groups of students: those assessed with paper-pencil and those assessed with a computer. All the computerized assessments were administered on a personal desktop or laptop computer. As required by the state, at the beginning of each computerized assessment, there were preparatory tutorial testing items that allowed students to familiarize themselves with the testing environment. Across the three academic years, the EOC biology data had a sample size of 33,401 high school students with 19,247 (57.6%) being assessed on a computer. Across the two academic years, the EOG ELA data included 23,939 eighth-grade students, 8,658 (36.2%) of whom were assessed on a computer. For the two academic years, the EOG math data consisted of 18,143 eighth-grade students, 7,256 (40%) of whom took the test using a computer. Student demographic descriptive statistics are shown in Table 1.

### Variables

In the dataset, a test-mode variable was included indicating whether students completed the assessment with paper-pencil or on a computer. The goal was to compare achievement scores of students in the paper-pencil assessment group with students in the computer assessment group. Achievement score values are presented in Table 1. For the high school EOC biology assessment, the paper-pencil group had a mean score of 252.80 ($SD$ = 10.42) and the computer group had a mean score of 251.97 ($SD$ = 10.75). For the eighth-grade EOG ELA assessment, the paper-pencil group had a mean score of 461.27 ($SD$ = 11.32) and the computer group had a mean score of 459.43 ($SD$ = 11.84). For the eighth-grade EOG math assessment, the paper-pencil group had a mean score of 451.52 ($SD$ = 10.99) and the computer group had a mean score of 447.75 ($SD$ = 10.18).

Prior academic achievement was also included as a variable in the dataset (Table 1). For high school EOC biology, eighth-grade EOG science was used as prior achievement (paper-pencil group: $M$ = 254.86, $SD$ = 10.23; computer group: $M$ = 254.60, $SD$ = 10.09). For eighth-grade EOG ELA, seventh-grade EOG ELA was used as prior achievement (paper-pencil group: $M$ =

Table 1
*Variable Descriptives by Subject Before Matching*

| | High School Biology | | | | Eighth Grade ELA | | | | Eighth Grade Math | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Paper-pencil | | Computer | | Paper-pencil | | Computer | | Paper-pencil | | Computer | |
| | M | SD | M | SD | M | SD | M | SD | M | SD | M | SD |
| Scale Score | 252.80 | 10.42 | 251.97 | 10.75 | 461.27 | 11.32 | 459.43 | 11.84 | 451.52 | 10.99 | 447.75 | 10.18 |
| Prior Achievement | 254.86 | 10.23 | 254.60 | 10.09 | 458.28 | 11.40 | 457.31 | 11.73 | 452.09 | 10.50 | 448.78 | 9.97 |
| Demographics | | | | | | | | | | | | |
| Female | 51.03% | | 49.56% | | 49.34% | | 48.30% | | 49.04% | | 47.53% | |
| American Indian or Alaska Native | 0.29% | | 0.32% | | 0.29% | | 0.29% | | 0.31% | | 0.34% | |
| Asian | 6.32% | | 8.26% | | 7.76% | | 9.25% | | 6.31% | | 6.96% | |
| Hispanic or Latino | 15.92% | | 15.53% | | 16.70% | | 18.72% | | 18.30% | | 23.35% | |
| Black or African American | 24.06% | | 26.88% | | 21.86% | | 25.47% | | 25.69% | | 28.97% | |
| White | 49.41% | | 44.75% | | 49.24% | | 42.22% | | 45.61% | | 36.25% | |
| Multiracial | 3.85% | | 4.08% | | 4.03% | | 3.98% | | 3.63% | | 4.08% | |
| Native Hawaiian or Other Pacific Islander | 0.14% | | 0.17% | | 0.12% | | 0.07% | | 0.15% | | 0.06% | |
| Limited English Proficiency | 4.37% | | 4.45% | | 5.05% | | 5.84% | | 5.73% | | 8.32% | |
| Special Education Needs | 12.06% | | 11.76% | | 12.64% | | 13.24% | | 15.32% | | 18.03% | |
| Academically Gifted | 25.88% | | 21.84% | | 25.19% | | 23.94% | | 19.97% | | 12.91% | |
| Assessment Year 2016 | 44.45% | | 34.06% | | NA | | NA | | NA | | NA | |
| Assessment Year 2017 | 36.27% | | 38.60% | | 56.42% | | 41.86% | | 79.00% | | 50.23% | |
| Assessment Year 2018 | 19.28% | | 27.33% | | 43.58% | | 58.14% | | 21.00% | | 49.77% | |
| ***N (Sample Size)*** | 14,154 | | 19,247 | | 15,281 | | 8,658 | | 10,887 | | 7,256 | |

*Note*. Scale Score is students' achievement in high school EOC biology, eighth-grade EOG ELA, and eighth-grade EOG math. Eighth-grade EOG science was used as high school EOC biology Prior Achievement. Seventh-grade EOG ELA was used as eighth-grade ELA Prior Achievement. Seventh-grade EOG math was used as eighth-grade EOG math Prior Achievement.

458.28, *SD* = 11.40; computer group: *M* = 457.31, *SD* = 11.73). For eighth-grade EOG math, seventh-grade EOG math was used as prior achievement (paper-pencil group: *M* = 452.09, *SD* = 10.50; computer group: *M* = 448.78, *SD* = 9.97).

Finally, student background variables including gender, race/ethnicity, English language proficiency, special education needs, academically gifted status, and year of the assessment were used. The detailed descriptive statistics for these demographics are provided in Table 1.

**Analytical Procedure**

The analysis for this study involved four major steps. In the first step, two-sample *t*-tests for each of the three subjects were conducted to explore any performance differences between the paper-pencil and computer format assessment groups. In the second step, nearest neighbor propensity score matching (PSM) was used to match students assessed with paper-pencil with students assessed on computer. PSM was chosen to increase meaningful comparisons between the two groups of students (for practical step-by-step guides on conducting PSM, refer to Ho et al. [2011] and Randolph et al. [2014]). Given that students were not randomly assigned to one of the two testing modes, a PSM approach helps reduce bias due to covariates (Austin, 2014; Randolph et al., 2014; Rosenbaum & Rubin, 1983). Matching was based on students' prior achievement, gender, race/ethnicity, English language proficiency, special education needs, academically gifted status, and year of the assessment. Matching was completed with the statistical software packages *R* and *MatchIt* (Ho et al. 2011; Randolph et al., 2014). Prior to matching, the following variables were dummy coded: gender, race/ethnicity, English language proficiency status, special education needs, and academically gifted status. Given that there were seven racial/ethnic groups, seven dummy variables for race/ethnicity were created. The descriptive statistics for the demographic variables after matching are provided in Table 2.

In the third step, for each of the three academic subjects, two-sample *t*-tests were run to compare the achievement scores of students across the two testing modes after PSM. To evaluate if any group mean differences between the two matched groups were meaningful, effect sizes were estimated. Furthermore, whether these differences were meaningful in relation to the state's established proficiency levels was also explored. To do so, for students in the computer assessment group, group mean differences in EOC and EOG achievement data were converted to percentage change in proficiency. If there was a one-point difference in test mode effect favoring students in the paper-pencil group, the number of students falling one point below the proficiency benchmark were calculated. The percentage change in proficiency was estimated by taking the number of students who scored one point below proficiency, dividing by the total number of test takers in the sample for each subject, and multiplying by 100. Converting to percentage change in proficiency demonstrates what percentage of students might move from below proficiency to proficiency if they were assessed with paper-pencil instead of on a computer.

In the last step, the PSM matched samples were analyzed with generalized linear models (GLM). The rationale of going beyond *t*-tests was that the PSM samples might still differ in student demographics and prior achievement. The GLM approach controls for these factors in order to

Table 2
*Variable Descritpives by Subject After Matching*

| | High School Biology | | | | Eighth Grade ELA | | | | Eighth Grade Math | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Paper-pencil | | Computer | | Paper-pencil | | Computer | | Paper-pencil | | Computer | |
| | M | SD | M | SD | M | SD | M | SD | M | SD | M | SD |
| Scale Score | 254.00 | 9.93 | 252.20 | 10.40 | 460.90 | 11.37 | 459.80 | 11.69 | 449.00 | 10.83 | 447.80 | 10.10 |
| Prior Achievement | 254.86 | 10.23 | 254.09 | 10.06 | 457.54 | 11.61 | 457.32 | 11.72 | 449.14 | 10.50 | 448.78 | 9.97 |
| Demographics | | | | | | | | | | | | |
| Female | 51.63% | | 45.16% | | 47.68% | | 48.44% | | 46.89% | | 47.57% | |
| American Indian or Alaska Native | 0.27% | | 0.47% | | 0.23% | | 0.31% | | 0.33% | | 0.37% | |
| Asian | 6.69% | | 10.70% | | 9.18% | | 8.93% | | 8.79% | | 6.60% | |
| Hispanic or Latino | 14.64% | | 14.32% | | 18.69% | | 18.69% | | 24.55% | | 23.57% | |
| Black or African American | 22.76% | | 29.58% | | 25.47% | | 25.63% | | 29.58% | | 29.33% | |
| White | 51.56% | | 40.10% | | 42.48% | | 42.42% | | 32.06% | | 36.03% | |
| Multiracial | 3.97% | | 4.65% | | 3.90% | | 3.95% | | 4.61% | | 4.04% | |
| Native Hawaiian or Other Pacific Islander | 0.11% | | 0.19% | | 0.05% | | 0.08% | | 0.08% | | 0.06% | |
| Limited English Proficiency | 2.52% | | 2.96% | | 3.99% | | 4.54% | | 7.04% | | 7.29% | |
| Special Education Needs | 10.03% | | 11.08% | | 12.63% | | 13.15% | | 18.16% | | 18.44% | |
| Academically Gifted | 30.50% | | 19.61% | | 25.95% | | 25.22% | | 15.99% | | 13.25% | |
| Assessment Year 2016 | 40.10% | | 10.05% | | NA | | NA | | NA | | NA | |
| Assessment Year 2017 | 37.97% | | 48.64% | | 39.86% | | 39.90% | | 68.09% | | 48.81% | |
| Assessment Year 2018 | 21.94% | | 41.31% | | 60.14% | | 60.10% | | 31.91% | | 51.19% | |
| ***N (Sample Size)*** | 11,323 | | 11,323 | | 7,828 | | 7,828 | | 6,437 | | 6,437 | |

provide a more robust estimate on the difference between the two groups of students in each subject. Equation (1) denotes the GLM utilized:

$$Y_{eoc/eog} = \beta_0 + \beta_1 x_{pc} + \beta_n X_n + \delta + \varepsilon \quad (1)$$

Y represents the dependent variable, student test score on high school EOC biology or eighth-grade EOG ELA or math. $\beta_0$ is the intercept which is the mean value of the dependent variable when the predictor variables in the model are zero. The $x_{pc}$ , $\beta_1$ as the coefficient, is the variable indicating whether students completed the assessment with paper-pencil or on a computer. The $X_n$, $\beta_n$ as coefficient, represents a set of control variables including prior achievement, gender, race/ethnicity, English language proficiency, special education needs, academically gifted status, and assessment year. The δ denotes the school fixed effects that handle school-level confounding factors/variables. The error term, $\varepsilon$, shows to what extent the observed data differ from what the model predicts.

## Results

Results from the two-sample *t*-tests before PSM are presented first followed by the two-sample *t*-tests after PSM. Next, group mean differences, their effect sizes, and proficiency achievement levels are examined. Finally, results from the GLM models are reported.

Descriptive statistics and *t*-test results before PSM are displayed in Tables 3.1 through 3.3. For high school EOC biology, in comparison to the computer group of students, the paper-pencil group of students had a significantly higher level of prior achievement as well as a higher percentage of academically gifted students. They also had a significantly higher percentage of female and White students, and a lower percentage of Asian and Black students. For eighth-grade EOG ELA, in comparison to the computer group of students, the paper-pencil group of students had a significantly higher level of prior achievement. They also had a significantly higher percentage of White students, and a lower percentage of Asian, Hispanic/Latino, and Black students. For eighth-grade EOG math, in comparison to the computer group of students, the paper-pencil group of students also had a significantly higher level of prior achievement, a higher percentage of academically gifted students, and a lower percentage of limited English proficiency and special education needs students. They also had a significantly higher percentage of female and White students, and a lower percentage of Hispanic/Latino and Black students. These differences in descriptive statistics justified a PSM procedure in order to improve the comparability between the paper-pencil and computer-based groups.

In terms of test mode comparability before PSM, across all three assessments, students in the paper-pencil group achieved statistically significantly higher scores than students in the computer group (high school EOC biology assessment, .80 points; eighth-grade EOG ELA, 1.90 points; eighth-grade EOG math, 3.70 points).

**Table 3.1**
*T-Test Results Before Matching: High School Biology*

| | Paper-pencil | | Computer | | | | | |
|---|---|---|---|---|---|---|---|---|
| | M | SD | M | SD | Diff. | *t* | *p-value* | *Cohen's D Effect Size* |
| Scale Score | 252.80 | 10.42 | 252.00 | 10.75 | -0.80 | 7.05 | < **.001** | -0.08 |
| Prior Achievement | 254.90 | 10.23 | 254.60 | 10.09 | -0.30 | 2.04 | **0.042** | -0.03 |
| Demographics | | | | | | | | |
| Female | 51.03% | | 49.56% | | -1.47% | 2.66 | **0.008** | |
| American Indian or Alaska Native | 0.29% | | 0.32% | | 0.03% | -0.53 | 0.597 | |
| Asian | 6.32% | | 8.26% | | 1.94% | -6.66 | < **. 001** | |
| Hispanic or Latino | 15.92% | | 15.53% | | -0.39% | 0.95 | 0.342 | |
| Black or African | 24.06% | | 26.88% | | 2.82% | -5.84 | < **.001** | |
| White | 49.41% | | 44.75% | | -4.66% | 8.45 | < **.001** | |
| Multiracial | 3.85% | | 4.08% | | 0.23% | -1.05 | 0.290 | |
| Native Hawaiian or Other Pacific Islander | 0.14% | | 0.17% | | 0.03% | -0.57 | 0.568 | |
| Limited English Proficiency | 4.37% | | 4.45% | | 0.08% | -0.36 | 0.721 | |
| Special Education Needs | 12.06% | | 11.76% | | -0.30% | 0.84 | 0.399 | |
| Academically Gifted | 25.88% | | 21.84% | | -4.04% | 8.60 | < **.001** | |
| Assessment Year 2016 | 44.45% | | 34.08% | | -10.37% | 19.39 | <**. 001** | |
| Assessment Year 2017 | 36.27% | | 38.60% | | 2.33% | -4.34 | < **.001** | |
| Assessment Year 2018 | 19.28% | | 27.33% | | 8.05% | -17.12 | < **.001** | |

**Table 3.2**
*T-Test Results Before Matching: Eighth Grade ELA*

| | Paper-pencil | | Computer | | | | | |
|---|---|---|---|---|---|---|---|---|
| | M | SD | M | SD | Diff. | *t* | *p-value* | *Cohen's D Effect Size* |
| Scale Score | 461.30 | 11.32 | 459.40 | 11.84 | -1.90 | 11.80 | < **.001** | -0.17 |
| Prior Achievement | 458.30 | 11.40 | 457.30 | 11.73 | -1.00 | 5.91 | < **.001** | -0.09 |
| Demographics | | | | | | | | |
| Female | 49.34% | | 48.30% | | -1.04% | 1.54 | 0.124 | |
| American Indian or Alaska Native | 0.29% | | 0.29% | | 0.00% | -0.01 | 0.991 | |
| Asian | 7.76% | | 9.25% | | 1.49% | -4.02 | < **.001** | |
| Hispanic or Latino | 16.70% | | 18.72% | | 2.02% | -3.96 | < **.001** | |
| Black or African American | 21.86% | | 25.47% | | 3.61% | -6.37 | < **.001** | |
| White | 49.24% | | 42.22% | | -7.02% | 10.50 | < **.001** | |
| Multiracial | 4.03% | | 3.98% | | -0.05% | 0.18 | 0.861 | |
| Native Hawaiian or Other Pacific Islander | 0.12% | | 0.07% | | -0.05% | 1.14 | 0.255 | |
| Limited English Proficiency | 5.05% | | 5.84% | | 0.79% | -2.64 | 0.008 | |
| Special Education Needs | 12.64% | | 13.24% | | 0.60% | -1.33 | 0.183 | |
| Academically Gifted | 25.19% | | 23.94% | | -1.25% | 2.15 | 0.032 | |
| Assessment Year 2016 | NA | | NA | | NA | NA | NA | |
| Assessment Year 2017 | 56.42% | | 41.86% | | -14.56% | 21.88 | < **.001** | |
| Assessment Year 2018 | 43.58% | | 58.14% | | 14.56% | -21.88 | < **.001** | |

**Table 3.3**
*T-Test Results Before Matching: Eighth Grade Math*

| | Paper-pencil | | Computer | | | | | |
|---|---|---|---|---|---|---|---|---|
| | M | SD | M | SD | Diff. | *t* | *p-value* | *Cohen's D Effect Size* |
| Scale Score | 451.50 | 10.99 | 447.80 | 10.18 | -3.70 | 23.19 | < **.001** | -0.35 |
| Prior Achievement | 452.10 | 10.50 | 448.80 | 9.97 | -3.30 | 19.97 | < **.001** | -0.32 |
| Demographics | | | | | | | | |
| Female | 49.04% | | 47.53% | | -1.51% | 1.99 | **0.047** | |
| American Indian or Alaska Native | 0.31% | | 0.35% | | 0.03% | -0.37 | 0.709 | |
| Asian | 6.31% | | 6.96% | | 0.65% | -1.73 | 0.084 | |
| Hispanic or Latino | 18.30% | | 23.35% | | 5.05% | -8.30 | < **.001** | |
| Black or African American | 25.69% | | 28.97% | | 3.28% | -4.87 | < **.001** | |
| White | 45.61% | | 36.25% | | -9.36% | 12.58 | < **.001** | |
| Multiracial | 3.63% | | 4.08% | | 0.45% | -1.56 | 0.120 | |
| Native Hawaiian or Other Pacific Islander | 0.15% | | 0.06% | | -0.09% | 1.83 | 0.068 | |
| Limited English Proficiency | 5.73% | | 8.32% | | 2.59% | -6.82 | < **.001** | |
| Special Education Needs | 15.32% | | 18.03% | | 2.71% | -4.82 | < **.001** | |
| Academically Gifted | 19.97% | | 12.91% | | -7.06% | 12.40 | < **.001** | |
| Assessment Year 2016 | NA | | NA | | NA | NA | NA | |
| Assessment Year 2017 | 79.00% | | 50.23% | | -28.77% | 42.50 | < **.001** | |
| Assessment Year 2018 | 21.00% | | 49.77% | | 28.77% | 42.50 | < **.001** | |

Descriptive statistics and *t*-test results after PSM are presented in Tables 4.1 through 4.3. For high school EOC biology, in comparison to the computer group of students, the paper-pencil group of students had significantly higher levels of prior achievement as well as a higher percentage of academically gifted, female, and White students. They also had a significantly lower percentage of Asian, Black, limited English proficiency, and special education needs students. For eighth-grade EOG ELA, results showed that the demographics were similar between the two groups of students. For eighth-grade EOG math, in comparison to the computer group of students, the paper-pencil group of students had a significantly higher percentage of Asian and academically gifted students, and a lower percentage of White students. Altogether, these findings indicate that the PSM procedure improved the comparability of the two groups of students in terms of EOG prior achievement, but not EOC prior achievement. For two of the three assessments, significant differences in demographics remained. Given these remaining differences, a GLM model was used to further examine test mode effect on student achievement.

The GLM model controlled for potential confounding factors such as students' demographics and prior achievement.

**Table 4.1**
*T-Test Results After Matching: High School Biology*

| | Paper-pencil | | Computer | | | | | |
|---|---|---|---|---|---|---|---|---|
| | M | SD | M | SD | Diff. | *t* | *p-value* | *Cohen's D Effect Size* |
| Scale Score | 254.00 | 9.93 | 252.20 | 10.40 | -1.80 | 13.45 | < **.001** | -0.18 |
| Prior Achievement | 254.86 | 10.23 | 254.09 | 10.06 | -0.77 | 5.75 | < **.001** | -0.08 |
| Demographics | | | | | | | | |
| Female | 51.63% | | 45.16% | | -6.46% | 9.75 | < **.001** | |
| American Indian or Alaska Native | 0.27% | | 0.47% | | 0.19% | -2.41 | **0.016** | |
| Asian | 6.69% | | 10.70% | | 4.01% | -10.74 | < **.001** | |
| Hispanic or Latino | 14.64% | | 14.32% | | -0.32% | 0.68 | 0.497 | |
| Black or African | 22.76% | | 29.58% | | 6.82% | -11.71 | < **.001** | |
| White | 51.56% | | 40.10% | | -11.45% | 17.41 | < **.001** | |
| Multiracial | 3.97% | | 4.65% | | 0.67% | -2.49 | **0.013** | |
| Native Hawaiian or Other Pacific Islander | 0.11% | | 0.19% | | 0.08% | -1.57 | 0.117 | |
| Limited English Proficiency | 2.52% | | 2.96% | | 0.44% | -2.04 | **0.042** | |
| Special Education Needs | 10.03% | | 11.08% | | 1.05% | -2.57 | < **.001** | |
| Academically Gifted | 30.50% | | 19.61% | | -10.89% | 19.06 | < **.001** | |
| Assessment Year 2016 | 40.10% | | 10.05% | | -30.05% | 55.60 | < **.001** | |
| Assessment Year 2017 | 37.97% | | 48.64% | | 10.67% | -16.29 | < **.001** | |
| Assessment Year 2018 | 21.94% | | 41.31% | | 19.38% | -32.05 | < **.001** | |

**Table 4.2**
*T-Test Results After Matching: Eighth Grade ELA*

| | Paper-pencil | | Computer | | | | | |
|---|---|---|---|---|---|---|---|---|
| | M | SD | M | SD | Diff. | *t* | *p-value* | *Cohen's D Effect Size* |
| Scale Score | 460.90 | 11.37 | 459.80 | 11.69 | -1.10 | 5.96 | < .001 | -0.10 |
| Prior Achievement | 457.54 | 11.61 | 457.32 | 11.72 | -0.22 | 1.17 | 0.243 | -0.02 |
| Demographics | | | | | | | | |
| Female | 47.68% | | 48.44% | | 0.77% | -0.96 | 0.337 | |
| American Indian or Alaska Native | 0.23% | | 0.31% | | 0.08% | -0.93 | 0.354 | |
| Asian | 9.18% | | 8.93% | | -0.26% | 0.56 | 0.578 | |
| Hispanic or Latino | 18.69% | | 18.69% | | 0.00% | 0.00 | 1.000 | |
| Black or African American | 25.47% | | 25.63% | | 0.15% | -0.22 | 0.826 | |
| White | 42.48% | | 42.42% | | -0.05% | 0.06 | 0.948 | |
| Multiracial | 3.90% | | 3.95% | | 0.05% | -0.16 | 0.869 | |
| Native Hawaiian or Other Pacific Islander | 0.05% | | 0.08% | | 0.03% | -0.63 | 0.527 | |
| Limited English Proficiency | 3.99% | | 4.54% | | 0.55% | -1.70 | 0.089 | |
| Special Education Needs | 12.63% | | 13.15% | | 0.51% | -0.95 | 0.340 | |
| Academically Gifted | 25.95% | | 25.22% | | -0.73% | 1.04 | 0.297 | |
| Assessment Year 2016 | NA | | NA | | NA | NA | NA | |
| Assessment Year 2017 | 39.86% | | 39.90% | | 0.04% | -0.05 | 0.961 | |
| Assessment Year 2018 | 60.14% | | 60.10% | | -0.04% | 0.05 | 0.989 | |

**Table 4.3**
*T-Test Results After Matching: Eighth Grade Math*

| | Paper-pencil | | Computer | | | | | |
|---|---|---|---|---|---|---|---|---|
| | M | SD | M | SD | Diff. | *t* | *p-value* | *Cohen's D Effect Size* |
| Scale Score | 449.00 | 10.83 | 447.80 | 10.10 | -1.20 | 6.48 | < **.001** | -0.11 |
| Prior Achievement | 449.14 | 10.50 | 448.78 | 9.97 | -0.36 | 1.99 | 0.047 | -0.04 |
| Demographics | | | | | | | | |
| Female | 46.89% | | 47.57% | | 0.68% | -0.78 | 0.437 | |
| American Indian or Alaska Native | 0.33% | | 0.37% | | 0.05% | -0.45 | 0.654 | |
| Asian | 8.79% | | 6.60% | | -2.19% | 4.67 | < **.001** | |
| Hispanic or Latino | 24.55% | | 23.57% | | -0.98% | 1.30 | 0.194 | |
| Black or African American | 29.58% | | 29.33% | | -0.25% | 0.31 | 0.757 | |
| White | 32.06% | | 36.03% | | 3.96% | -4.75 | < **.001** | |
| Multiracial | 4.61% | | 4.04% | | -0.57% | 1.60 | 0.109 | |
| Native Hawaiian or Other Pacific Islander | 0.08% | | 0.06% | | -0.02% | 0.33 | 0.739 | |
| Limited English Proficiency | 7.04% | | 7.29% | | 0.25% | -0.55 | 0.585 | |
| Special Education Needs | 18.16% | | 18.44% | | 0.28% | -0.41 | 0.682 | |
| Academically Gifted | 15.99% | | 13.25% | | -2.73% | 4.39 | < **.001** | |
| Assessment Year 2016 | NA | | NA | | NA | NA | NA | |
| Assessment Year 2017 | 68.09% | | 48.81% | | -19.28% | 22.63 | < **.001** | |
| Assessment Year 2018 | 31.91% | | 51.19% | | 19.28% | -22.63 | < **.001** | |

The after-matching comparison of scale scores and summary of effect sizes are displayed in Table 5. Across all three assessments, students in the paper-pencil group achieved statistically significantly higher scores than students in the computer group (high school EOC biology assessment, 1.80 points; eighth-grade EOG ELA, 1.10 points; eighth-grade EOG math, 1.20 points). While test mode differences were shown across the three subjects, the subsequent question posed was if the one to two points score differences on the assessments were in fact meaningful. According to Cohen's D effect sizes (less than .2 denotes a small effect size), for all three assessments, the achievement differences between the two format groups were small (EOC biology = .18; EOG ELA = .10; EOG math = .11). It should be taken into account, however, that within the educational domain effect sizes are unlikely to approach .30 (Lipsey et al., 2012).

In terms of whether these test mode differences were meaningful in relation to achieving proficiency, the EOC biology cutoff score from non-grade level proficient to grade-level proficient was a score of 250 points. The 330 students (1.46%) in the computer group who

**Table 5**
*Summary of Effect Sizes by Subject After Matching*

| | Paper-pencil | | Computer | | | | | |
|---|---|---|---|---|---|---|---|---|
| | M | SD | M | SD | Diff. | *t* | *p-value* | *Cohen's D Effect Size* |
| High School Biology (N = 22,646) | 254.00 | 9.93 | 252.20 | 10.40 | -1.80 | 13.45 | < .001 | -0.18 |
| Eighth Grade ELA (N =15,656) | 460.90 | 11.37 | 459.80 | 11.69 | -1.10 | 5.96 | < .001 | -0.10 |
| Eighth Grade Math (N = 12,874) | 449.00 | 10.83 | 447.80 | 10.10 | -1.20 | 6.48 | < .001 | -0.11 |

scored 249 (just one point below the cutoff) could potentially have scored 250 points if they had taken the assessment with paper-pencil thereby earning grade-level proficiency. The EOG ELA and math cutoff scores from non-grade level proficient to grade-level proficient was a score of 458 and 452 points, respectively. Results suggest that 267 (1.71%) and 239 (1.86%) students in the computer group may have switched from non-proficient to grade-level proficient achievement had they taken the paper-pencil assessment. These results indicate that only a relatively small percentage of students may have benefitted from completing the paper-based format of the assessment. This suggests that completing the assessment using a computer, as opposed to paper-pencil, did not negatively impact the majority of students' performance on any of the three subjects. However, school districts view students as unique individuals and place importance on each student's right to learn, grow, and succeed. From that perspective, these results demonstrate that 836 individual students' educational paths (e.g., grade retention and course placement) could have been potentially impacted because they were instructed to complete an assessment using a computer as opposed to paper-pencil.

The GLM results with standardized coefficients are presented in Table 6. Among the control variables, prior achievement, race/ethnicity, English language proficiency, special education needs, and academically gifted status were associated with students' performance in all three subjects. For example, one *SD* increase in prior achievement was associated with a .69 *SD* increase in EOC biology. The Black and White student achievement gap was -.08 *SD* for EOC biology and -.06 *SD* for both EOG ELA and math with White students having higher achievement than Black students. The Hispanic and White student achievement gap was -.05 *SD* for EOC biology and -.04 *SD* for both EOG ELA and math with White students having higher achievement than Hispanic students. In terms of achievement and test mode comparability, for EOC biology, the computerized assessment was associated with a .05 *SD* decrease in achievement score. For EOG ELA and math, the computerized assessment was associated with a .07 and .03 *SD* decrease in achievement score, respectively. These results confirm the small but statistically significant difference found in the previously reviewed *t*-tests.

**Table 6**
*GLM Results of the Matched Samples by Subject*

| | High School Biology | Eighth Grade ELA | Eighth Grade Math |
|---|---|---|---|
| Intercept | 0.00 *** | 0.00 *** | 0.00 *** |
| | (1.19) | (2.71) | (2.85) |
| Prior Achievement | 0.69 *** | 0.70 *** | 0.71 *** |
| | (0.00) | (0.01) | (0.01) |
| Computerized | -0.05 *** | -0.07 *** | -0.03 *** |
| | (0.15) | (0.20) | (0.16) |
| Female | 0.04 *** | 0.05 *** | 0.00 |
| | (0.07) | (0.10) | (0.09) |
| Amer. Indian | -0.01 ** | -0.01 * | -0.01 |
| | (0.57) | (0.91) | (0.74) |
| Asian | 0.02 *** | 0.00 | 0.04 *** |
| | (0.14) | (0.18) | (0.19) |
| Hisp./Latino | -0.05 *** | -0.04 *** | -0.04 *** |
| | (0.11) | (0.15) | (0.13) |
| Black | -0.08 *** | -0.06 *** | -0.06 *** |
| | (0.10) | (0.13) | (0.12) |
| Multiracial | -0.02 *** | -0.01 ** | -0.01 ** |
| | (0.18) | (0.25) | (0.22) |
| Pacific Islander | 0.00 | 0.00 | 0.00 |
| | (0.91) | (1.86) | (1.64) |
| LEP | -0.03 *** | -0.03 *** | -0.01 ** |
| | (0.23) | (0.26) | (0.19) |
| SPED | -0.06 *** | -0.08 *** | -0.05 *** |
| | (0.12) | (0.16) | (0.12) |
| AIG | 0.09 *** | 0.10 *** | 0.11 *** |
| | (0.10) | (0.13) | (0.16) |
| Year 2016 | 0.04 *** | | |
| | (0.11) | | |
| Year 2017 | -0.03 *** | -0.03 *** | 0.03 *** |
| | (0.08) | (0.10) | (0.11) |

*Note.* * $p < .05$, ** $p < .01$, *** $p < .001$.

## Discussion

School districts are increasingly transitioning from traditional paper-and-pencil assessments to computer-based assessments. Computerized assessments offer several advantages which include, but are not limited to flexibility in designing test items, efficient administration, immediate scoring and reporting of results, reduced testing costs, increased student motivation and engagement, reduced opportunities for student and staff cheating, and consistency across

classroom activities and assessments that support students' computer literacy skills (Backes & Cowan, 2019; Bennett, 2003; Boo & Vispoel, 2012; EdTech Strategies, 2015; Ghaderi et al., 2014; Kim & Huynh, 2010; Randall et al., 2012; Thurlow et al., 2010; U.S. DoE, 2013). These benefits are highly compelling. Yet given that critical school decisions are made based on students' standardized test performance, it is recommended that districts not adopt online testing without careful preparation, conducting their own evaluations of test mode comparability, and exercising caution when interpreting and using transition-year scores. Doing so will help ensure that bias is not introduced into students' performance due to factors independent of their academic knowledge and skill levels.

The current investigation utilized data from a large urban school district transitioning to computerized testing and compared students' performance between paper-pencil and computer-based formats of large-scale statewide EOC and EOG assessments in biology, ELA, and math. Results revealed a testing mode effect across all three assessments: students who took the paper-pencil-based exams performed higher than students who took the computer-based exams. Based on Cohen's widely accepted classification of effect sizes, the magnitude of these differences was small. These findings are in alignment with the growing literature that has found small in magnitude, but significant test mode differences (e.g., Backes & Cowan, 2019; Bennett et al., 2008; Kim & Huynh, 2008). For example, Kim and Huynh (2007; 2008) found that on large-scale statewide EOC algebra and English exams, students performed higher on the paper-based format in comparison to the computer-based format; however, the effect sizes were classified as small. Also, in Bennett et al.'s (2008) investigation of eighth-grade students' performance on either a computer- or paper-based math test, results showed that students scored higher on the paper-based math items in comparison to the computer-based items, though these mode differences in general were small.

Although the current study's results consistently indicated that having students take the assessment using a computer was associated with a small decrease in test performance, these results should not be interpreted as negligible. Even though Cohen's classification of effect sizes is widely used, it has not been shown to be applicable in the educational domain where effect sizes are unlikely to approach .30 (Lipsey et al., 2012). It may in fact be misleading to apply Cohen's interpretation to the smaller effect sizes found within this study. Therefore, when the effect sizes found within this study are taken into consideration with respect to educational settings (EOC biology = .18; EOG ELA = .10; EOG math = .11), the results may indicate test mode differences. Furthermore, after including prior achievement as a control variable in the GLM models, for high school EOC biology, the computerization coefficient was the same as the Hispanic and White student achievement gap (both -.05 at $p < .001$). For eighth-grade EOG ELA, the computerization coefficient was similar to the Black and White student achievement gap (-.07 vs. -.06 at $p < .001$). These results demonstrate that the computerization test mode effect on achievement was comparable to certain racial achievement gaps, thereby indicating that computerization was associated with a meaningful decrease in performance scores. Moreover, the current school district's vision and mission are centered on the core belief that each student is unique and deserves to learn, grow, and succeed. Results from this investigation should be used to inform the district, as well as districts at large operating with similar beliefs, that 836 individual students' educational trajectories could have been altered or penalized as a result of

adapting to a new testing format. As districts move forward with transitioning to online testing, they should interpret potential transition-year test mode differences with caution.

One limitation of the current study is that students were not randomly assigned to either of the testing format groups. It may be the case that schools with higher technological capacity and infrastructure were more likely to opt-in for the computerized testing format thereby introducing sampling bias and limiting generalizability of findings across academic settings. To address this methodological drawback, PSM and GLM were used. In doing so, results contribute to the existing test mode effect literature by demonstrating a rigorous statistics approach to minimize bias in comparing the two groups of students' test results while controlling for student prior achievement, demographics, and school fixed effects. Nevertheless, given that this is not a randomized experimental design, it cannot be concluded that computerization caused the small decrease in student achievement. To show a causal effect of computerization in student assessment, future studies implementing an experimental design with random samples are needed.

Another limitation of the study was the inability to determine what specific factors contributed to the small achievement differences found in the sample. Computerization of student assessments involves several factors that could potentially impact how students perform including presentation characteristics, response requirements, general administration characteristics, and students' accessibility to and experience using technology (Bennett, 2003; Duque, 2016; Leeson, 2006). Among these characteristics, the latter is particularly noteworthy. If students have varying degrees of familiarity with technology, their experience interacting with the exam and recording responses may differ. Consequently, computerized tests may measure students' proficiency in computer literacy and not subject mastery. In response to this issue and in preparation for administering tests online, schools may consider providing practice sessions so students could familiarize themselves with the computer-assessment environment. If students experience difficulty, testing coordinators or teachers could intervene with guided practice. In doing so, this would help eliminate the test mode effect due to limited computer literacy.

Despite the limitations, this study makes two valuable contributions to the existing literature. First, it applied a rigorous statistical approach to minimize bias in comparing the two groups of students' test results based on recent testing data from a large urban school district. Second, it provides results that suggest a small test mode effect across standardized assessments of biology, ELA, and math such that students who completed the paper-pencil format had a small advantage over students who completed the computer-based format. Given the growing trend to computerization of student assessments, this study is germane to school districts transitioning to computerized assessments and assists them in investigating test mode comparability.

**Author Notes**

**Meghan B. Scrimgeour, Ph.D.** is in the Data, Research, and Accountability Department for the Wake County Public School System in Cary, NC. She is a research analyst and conducts evaluations of district programs which impact students or staff.

**Haigen H. Huang, Ph.D.** is in the Data, Research, and Accountability Department for the Wake County Public School System in Cary, NC. He is an educational researcher serving schools and local communities. His primary research interest focuses on policy issues related to educational equity.

Correspondence concerning this article should be addressed to Meghan Scrimgeour at mscrimgeour@wcpss.net

## References

American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessments (APA). (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: American Psychological Association.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics In Medicine, 33*(6), 1057-1069. https://doi.org/10.1002/sim.6004

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*, 89-103. https://doi.org/10.1016/j.econedurev.2018.12.007

Bennett, R. E. (2003). *Online assessment and the comparability of score meaning* (RM-03-05). Princeton, NJ: Educational Testing Service.

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 on mode effects in NAEP. *The Journal of Technology, Learning, and Assessment, 6*(9), 1-39.

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

Choi, S. W., & Tinkler, T. (2002, April). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting.* Paper session presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Davis, M. R. (2014). Online testing glitches causing distrust in technology. *Education Week, 33*(30), 20-21.

Davis, L. L., Kong, X., McBride, Y., & Morrison, K. M. (2017). Device comparability of tablets and computers for assessment purposes. *Applied Measurement in Education, 30*(1), 16-26. https://doi.org/10.1080/08957347.2016.1243538

DePascale, C., Dadey, N., & Lyons, S. (2016). *Score comparability across computerized assessment delivery devices: Defining comparability, reviewing the literature, and providing recommendations for states when submitting to Title 1 peer review*. Washington, D.C.: Council of Chief State School Officers.

Duque, M. (2016). *Is there a PARCC mode effect?* (SDP Fellowship Capstone Report 2016). Harvard University, Center for Education Policy Research.

EdTech Strategies. (2015). *Pencils down: The shift to online and computer-based testing*. Retrieved from https://www.edtechstrategies.com/wp-content/uploads/2015/11/PencilsDownK-8_EdTech-StrategiesLLC.pdf

Farmer, B. (2016, February 20). *The state that pulled the plug on computer testing*. NPR.

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

Ho, D. E., Imai, K, King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*(8), 1-29.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B. A., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *The Journal of Technology, Learning and Assessment*, *5*(2), 1-50.

International Test Commission (2005). *International guidelines on computer-based and internet delivered testing*. Retrieved from https://www.intestcom.org/files/guideline_computer_based_testing.pdf

Kim, D., & Huynh, H. (2007). Comparability of computer and pencil-and-paper versions of algebra and biology assessments. *The Journal of Technology, Learning, and Assessment*, *6*(4), 1-31.

Kim, D. H., & Huynh, H. (2008). Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course English test. *Educational and Psychological Measurement*, *68*(4), 554-570.

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.

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.

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.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. (NCSER 2013-

3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

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

Office of Assessment, Evaluation, and Research Services, University of North Carolina at Greensboro. (In progress). *Mode comparability analysis for North Carolina ELA7 end-of-course assessment*.

Office of Assessment and Information Services, Oregon Department of Education. (2007). *Comparability of student scores obtained from paper and computer administrations*.

Paek, P. L. (2005). Recent trends in comparability studies using testing and assessment to promote learning. *Pearson Educational Measurement*.

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.

Pomplun, M., Ritchie, T., & Custer, M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment, 11*(2), 127-143.

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.

Randolph, J. J., Falbe, K., Manuel A. K., & Balloun, J. L. (2014). A step-by-step guide to propensity score matching in R. *Practical Assessment, Research & Evaluation, 19*(18), 1-6.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(2), 1-47.

Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). *Computer-based testing: Practices and considerations* (Synthesis Report 78). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics (2013). *Testing integrity symposium: Issues and recommendations for best practice. February 28, 2012.* Retrieved from https://ies.ed.gov/whatsnew/conferences/?id=966&cid=2

Wang, S. (2004). *Online or paper: Does delivery affect results. Administration Mode Comparability for Stanford Diagnostic Reading and Mathematics Tests*. San Antonio, TX: Pearson Inc.

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

Wang, 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., Lin, C. H., & Kong, J. (2008, March). *Maintaining score equivalence as tests transition online: Issues, approaches and trends*. Paper session presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

White, S., Kim, Y. Y., Chen, J., & Liu, F. (2012). *Performance of fourth-grade students in the 2012 NAEP computer-based writing pilot assessment. Scores, text length, and use of editing tools* (NCES 2015-119). Washington, D.C.: U.S. Department of Education.