Comparability of Computer-Based and Paper-Based Science Assessments

Cari F. Herrmann-Abell, Joseph Hardcastle, and George E. DeBoer
AAAS Project 2061

Paper presented at the
2018 NARST Annual International Conference
Atlanta, GA
March 10-13, 2018

Abstract

We compared student's performance on a paper-based test (PBT) and three computer-based tests (CBTs). The three computer-based tests used different test navigation and answer selection features, allowing us to examine how these features affect student performance. The study sample consisted of 9,698 fourth through twelfth grade students from across the U.S. who were randomly assigned to take a test in one of the four modes. CBT modes differed in whether students could skip questions and freely move through the test, and whether students could click directly on the answer choice or had to click on a radio button at the bottom of the screen. Rasch analysis was used to estimate item difficulties and student performance levels. Student performance level was then used as an outcome in hiearchal linear models to determine the mode effects. We found that student performance was unaffected by whether the test was paper-based or computer-based. A comparison of student performance on the three CBTs indicated that restricting test navigation did not affect student performance, but allowing students to select an answer choice by directly clicking on it improved student performance. Our findings show that CBTs can be considered equivalent to PBTs, and the results can also be used to inform best practices for the design of other CBTs.

Introduction

With the increased availably of computers, many assessments are being administered as computer-based tests (CBT). CBT provides several advantages over paper-based tests (PBT) including ease and flexibility of administering and grading tests, as well as allowing for the development of novel technology-based testing environments (DeBoer et al., 2014). These benefits have made CBT increasingly popular; however, questions still remain about whether CBT- and PBT-generated scores can be considered equivalent measures of student performance. Despite the advantages of CBT over PBT, PBT are still frequently used by teachers due to limited access to computers. To address the fact that a test may be given in both CBT and PBT modalities, test developers must ensure that the CBT and PBT versions of a test can be considered equivalent.

Several studies have compared PBT and CBT (Leeson, 2006; Paek, 2005). Some have found little to no difference between them (e.g., Welch, 2014; Wang, Jiao, Young, Brooks, & Olson, 2008; Bridgeman, Lennon, & Jackenthal, 2003; Choi & Tinkler, 2002; Hetter, Segall, & Bloxom, 1994). Others have found student performance to be lower on CBT relative to PBT. These differences in student performance have been linked to technological differences, such as whether scrolling is required (Bridgeman et al., 2003; Choi & Tinkler, 2002), and to participant characteristics, such as the students' ethnicity, gender, or primary language (Gallagher, Bridgeman, & Cahalan, 2000). A synthesis of 81 studies also showed that subject had some impact on comparability, for example, CBT provided an advantage for English Language Arts tests, while PBT provided an advantage for Mathematics tests (Kingston, 2009). The research on the comparability of CBT and PBT provides some guidance for what to avoid when creating tests that will be administered in both modalities, but there is still a need to improve our understanding of best practices for the design and administration of equivalent CBT and PBT.

We are currently working on a project to develop and validate a set of three vertically-equated assessment instruments to monitor how students' understanding of energy progresses from late elementary to high school. Part of the validation process involves confirming that different testing modes result in equivalent student measures. In a preliminary study, we used a quasi-experimental design to test the comparability of the PBT version and two CBT versions of our multiple-choice assessment instrument (Hardcastle, Herrmann-Abell, & DeBoer, 2017). Propensity score matching was used to create equivalent demographic groups for each testing modality. Our analysis suggested that that elementary and middle school students performed worse on a computer-based test that did not allow them to return to previous items and that required them to select their answer choice at the bottom of the page, instead of directly clicking on the answer choice. However, the study design did not allow us to separate these variables. The study described here involved a randomized control trial to more rigorously investigate how test modality and specific test features influence student performance on the same test questions.

In this study, students were randomly assigned to take a test in one of four test modes—one paper-based and three different computer-based modes. Each computer-based test differed in whether it allowed the student to skip questions and freely move through the test, and whether students clicked directly on the answer choice or on a radio button at the bottom of the screen. Students' performances on different testing modalities were compared to evaluate whether PBT and CBT yielded equivalent measures of student knowledge. Students' performances on the three different CBT versions were examined to evaluate the effect of specific test navigation and answer choice selection features.

Methodology

Participants

A total of 10,779 students in the fourth through twelfth grades from 38 states and Puerto Rico participated in the study. Students who answered fewer than six of 35 items and students who did not provide all the requested demographic information were excluded from the study. Table 1 presents a summary of the demographic information broken down by testing modality for the 9,698 students included in the study. All students were enrolled in a science class at the time of testing, but not necessarily in a physical science class. Students within a classroom were randomly assigned to either the paper version or one of three computer-based versions. They were given one class period to complete as many items on the test as they could.

Table 1: *Summary of student demographics by testing modality*

	CBT 1 (n = 2255)	CBT 2 (n = 2249)	CBT 3 (n = 2195)	PBT (n = 2999)
Grade Band				
Elementary (4th-5th)	15%	14%	14%	14%
Middle (6th-8th)	42%	43%	43%	43%
High (9th-12th)	43%	43%	43%	43%
Gender				
Female	55%	55%	54%	54%
Male	45%	45%	46%	46%
Race/Ethnicity				
White	58%	58%	58%	57%
Hispanic	15%	14%	14%	13%
Black	12%	12%	11%	12%
Asian	5%	4%	6%	4%
American Indian	1%	2%	2%	2%
Pacific Islander	1%	1%	1%	0%
Other	8%	10%	9%	12%
Primary Language				
English	94%	93%	94%	91%
Other	6%	7%	6%	9%

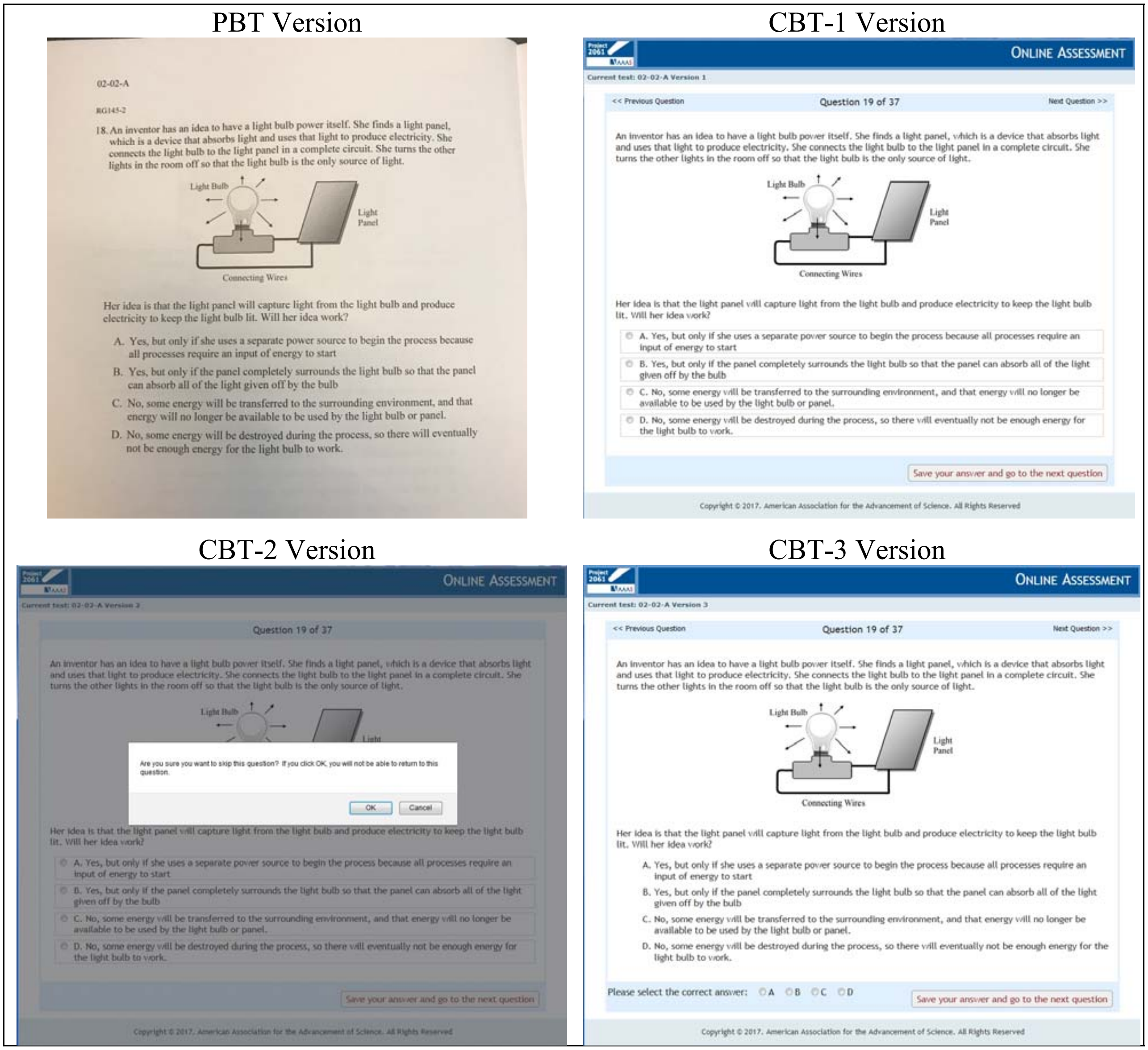
Assessments

The assessment instruments used during this study were developed as part of larger project aimed at assessing students understanding of energy from fourth through twelfth grades (Herrmann-Abell & DeBoer, 2018). The items were distractor-driven, multiple-choice items that were scored dichotomously. Items assessed students' understanding of (1) energy forms and transformations, (2) energy transfer, (3) energy dissipation and degradation, and (4) energy conservation. Three instruments, each consisting of 35 items, were used: two appropriate for all students, and one that targeted more advanced ideas, which was administered only to middle and high school students. Linking items were used so that item and student measures could be compared across instruments.

Testing modalities

We investigated four different testing modalities: one paper-and-pencil and three computer-based tests. Table 2 summarizes the characteristics of the different testing modalities and screen shots are shown in Figure 1. All tests used the same items with identical text, images, and answer choices. Students who took the PBT version were given an 8.5x11 test booklet and a scannable answer sheet. The test booklet was printed in black and white and used a serif font. Serif font has been shown to provide good readability for print media (Mohamad Ali, Wahid, Samsudin, & Zaffwan Idris, 2013).

The CBT versions were administered using our test generation and administration website (AAAS, n.d.). About 73% of the students used a laptop during the testing, 22% used a desktop, 5% used a tablet, and less than 1% used a smart phone. All three CBT versions included black and white images and used a sans-serif font. Previous studies have found no statistical difference in readability when comparing serif and sans-serif fonts on computer displays (Arditi & Cho, 2005; Mohamad Ali, et al., 2013); however, we chose a sans-serif font because younger children have indicated a preference for it when reading on a computer screen (Bernard, Chaparro, Mills, & Halcomb, 2002).



PBT Version

02-02-A

RG145-2

18. An inventor has an idea to have a light bulb power itself. She finds a light panel, which is a device that absorbs light and uses that light to produce electricity. She connects the light bulb to the light panel in a complete circuit. She turns the other lights in the room off so that the light bulb is the only source of light.

Her idea is that the light panel will capture light from the light bulb and produce electricity to keep the light bulb lit. Will her idea work?

A. Yes, but only if she uses a separate power source to begin the process because all processes require an input of energy to start

B. Yes, but only if the panel completely surrounds the light bulb so that the panel can absorb all of the light given off by the bulb

C. No, some energy will be transferred to the surrounding environment, and that energy will no longer be available to be used by the light bulb or panel.

D. No, some energy will be destroyed during the process, so there will eventually not be enough energy for the light bulb to work.

CBT-1 Version

Current test: 02-02-A Version 1

An inventor has an idea to have a light bulb power itself. She finds a light panel, which is a device that absorbs light and uses that light to produce electricity. She connects the light bulb to the light panel in a complete circuit. She turns the other lights in the room off so that the light bulb is the only source of light.

Her idea is that the light panel will capture light from the light bulb and produce electricity to keep the light bulb lit. Will her idea work?

○ A. Yes, but only if she uses a separate power source to begin the process because all processes require an input of energy to start

○ B. Yes, but only if the panel completely surrounds the light bulb so that the panel can absorb all of the light given off by the bulb

○ C. No, some energy will be transferred to the surrounding environment, and that energy will no longer be available to be used by the light bulb or panel.

○ D. No, some energy will be destroyed during the process, so there will eventually not be enough energy for the light bulb to work.

Copyright © 2017. American Association for the Advancement of Science. All Rights Reserved

CBT-2 Version

Current test: 02-02-A Version 2

An inventor has an idea to have a light bulb power itself. She finds a light panel, which is a device that absorbs light and uses that light to produce electricity. She connects the light bulb to the light panel in a complete circuit. She turns the other lights in the room off so that the light bulb is the only source of light.

Are you sure you want to skip this question? If you click OK, you will not be able to return to this question.

Her idea is that the light panel will capture light from the light bulb and produce electricity to keep the light bulb lit. Will her idea work?

○ A. Yes, but only if she uses a separate power source to begin the process because all processes require an input of energy to start

○ B. Yes, but only if the panel completely surrounds the light bulb so that the panel can absorb all of the light given off by the bulb

○ C. No, some energy will be transferred to the surrounding environment, and that energy will no longer be available to be used by the light bulb or panel.

○ D. No, some energy will be destroyed during the process, so there will eventually not be enough energy for the light bulb to work.

Copyright © 2017. American Association for the Advancement of Science. All Rights Reserved

CBT-3 Version

Current test: 02-02-A Version 3

An inventor has an idea to have a light bulb power itself. She finds a light panel, which is a device that absorbs light and uses that light to produce electricity. She connects the light bulb to the light panel in a complete circuit. She turns the other lights in the room off so that the light bulb is the only source of light.

Her idea is that the light panel will capture light from the light bulb and produce electricity to keep the light bulb lit. Will her idea work?

A. Yes, but only if she uses a separate power source to begin the process because all processes require an input of energy to start

B. Yes, but only if the panel completely surrounds the light bulb so that the panel can absorb all of the light given off by the bulb

C. No, some energy will be transferred to the surrounding environment, and that energy will no longer be available to be used by the light bulb or panel.

D. No, some energy will be destroyed during the process, so there will eventually not be enough energy for the light bulb to work.

Please select the correct answer: ○ A ○ B ○ C ○ D

Copyright © 2017. American Association for the Advancement of Science. All Rights Reserved

Figure 1: Screen shots of the different testing modalities. The screen shot of the CBT-2 version shows the warning message students receive if they try to skip a question.

The differences among the CBT versions include the way students selected an answer choice and how students navigated between items. In the first version (CBT-1), students could skip items and return to previous items, and they had to click directly on an answer choice to select it. In the second version (CBT- 2), students could skip items but could not return to previous items, and they had to click directly on an answer choice to select it. In the third version (CBT-3), students could skip items and return to previous items, but they could not directly click on answer choices. In the CBT-3 version, students clicked on “radio” buttons corresponding to their answer choice, which were located underneath the item. The no-skip, no-direct select condition was not tested.

Table 2: *Characteristics of the different testing modalities*

	PBT Version	CBT Version 1	CBT Version 2	CBT Version 3
Font	Serif	Sans-Serif	Sans-Serif	Sans-Serif
Images	Black & White	Black & White	Black & White	Black & White
Test Navigation	Students could skip items and freely move through test	Students could skip items and freely move through test	Students could not return to items once skipped	Students could skip items and freely move through test
Answer Choice Selection	Students "bubbled" in the letter of their answer choice on a separate sheet	Students clicked directly on their answer choice	Students clicked directly on their answer choice	Students clicked a "radio" button corresponding to their answer choice

Rasch Analysis

WINSTEPS software (Linacre, 2016) was used to estimate Rasch student and item measures. In the dichotomous Rasch model, the probability that a student will respond to an item correctly is determined by the difference in the student's performance level and the item's difficulty (Bond & Fox, 2007). The measures are expressed on the same interval scale, are measured in logits, and are mutually independent. The average item difficulty was set to zero logits.

The data's fit to the Rasch model was evaluated using separation indices, infit and outfit mean-squares, standard errors, and point-measure correlations. In an initial analysis, the fit statistics showed that there were two items with outfit mean-square values outside of the acceptable range of 0.7-1.3 (Bond & Fox, 2007). The outfit statistic was used because it is unweighted and, therefore, sensitive to outliers. The fit statistics for these two items suggested that some students were unexpectedly getting those items correct, possibly due to guessing.

One technique for addressing student guessing is to assume guessing is a function of the difficulty of the item and the proficiency of the student. If a low proficiency student gets a high difficulty item correct, it could be a sign that the student correctly guessed. In the Rasch model, this information is captured in students z-residuals for an item, where a high z-residual indicates the student was unlikely to answer the item correctly but actually did. To decrease the influence of guessing on our measures we used an approach outlined by Andrich *et al.* (2012), in which a tailored data set is created by replacing student responses that have large z-residual values with missing data. We replaced all student responses with z-residuals greater than 2.58 with missing data resulting in a total of 4,253 responses being replaced (approximately 1% of the total data). After removing these students' responses, three students were removed from the data set because we now had fewer than six responses from them.

The revised data set was modeled and Table 3 presents the final fit statistics for both items sand students. Overall, the data has an acceptable fit to the model. The separation index for the *items* was 22.03, which corresponds to a reliability of 1.00. The separation index for the *students* was

1.98, which corresponds to a reliability of 0.80. All the items had infit and outfit mean-square values within the acceptable range, and the point-measure correlations were all positive.

Table 3: *Summary of Rasch Fit Statistics*

	Item			Student		
	Min	Max	Median	Min	Max	Median
Standard error	0.03	0.07	0.04	0.36	1.90	0.39
Infit mean-square	0.82	1.29	0.99	0.44	1.99	0.99
Outfit mean-square	0.60	1.31	0.97	0.18	2.35	0.96
Point-measure correlation coefficients	0.10	0.55	0.35	-0.66	0.94	0.34
Separation index (Reliability)	22.03 (1.00)			1.98 (0.80)		

Hierarchal Linear Modeling

The Rasch student measures were modeled as the outcome in two-level hierarchical linear models (HLMs) with students at level 1 and teachers at level 2. Student-level variables included gender, race/ethnicity, and primary language. There were no teacher-level variables included.

A fully unconditional model containing only the outcome variable and no independent variables, except an intercept, was estimated first. This was followed by a conditional model in which gender, primary language, and race/ethnicity were included as controls and modeled as fixed effects. Then similar models were conducted on subsets of the data including models for each grade band and a model for students whose primary language was not English. HLM 7 software was used in this study (Raudenbush et al., 2011). The method of estimation was restricted maximum likelihood. Effect sizes were calculated by dividing the coefficient of the modality by the square root of the pooled student-level unadjusted SD.

Results

Fully Unconditional HLM

A fully unconditional HLM with no independent variables at either level was run to calculate the intraclass correlation coefficient. The results of the model are shown in Table 4. The intraclass correlation coefficient represents the proportion of variance in student performance level that could be the result of class characteristics. In this case, approximately 27% of the variance in student performance could be the function of class characteristics. Therefore, the proportion of the variance in student performance that exists at the individual level is 73%. A chi-square test indicated that posttest scores varied significantly between classes ($\chi^2 = 3060$, $p < 0.001$).

Table 4: *Fully unconditional HLM*

Variable	Value
Within-classroom variance (σ^2)	0.98
Between-classroom variance (τ)	0.36
Between-classroom SD	0.99
Reliability (λ)	0.94
Intraclass correlation (ρ)	0.27

Conditional HLM of Entire Student Sample

The mixed-model for the conditional HLM of the full data set was

$$\text{MEASURE}_{ij} = \gamma_{00} + \gamma_{10}{*}\text{CBT1}_{ij} + \gamma_{20}{*}\text{CBT2}_{ij} + \gamma_{30}{*}\text{CBT3}_{ij} + \gamma_{40}{*}\text{FEMALE}_{ij} + \gamma_{50}{*}\text{HI}_{ij} + \gamma_{60}{*}\text{BL}_{ij} + \gamma_{70}{*}\text{AS}_{ij} + \gamma_{80}{*}\text{AI}_{ij} + \gamma_{90}{*}\text{PI}_{ij} + \gamma_{100}{*}\text{OT}_{ij} + \gamma_{110}{*}\text{ENGLISH}_{ij} + u_{0j}+ r_{ij}$$

where MEASURE_{ij} is the Rasch student performance level for student *i* within teacher *j*. Three dummy variables, CBT1, CBT2, and CBT3, were created for the each of the computer-based test modalities. The students who took the paper-based test modality were used as a reference group. FEMALE is a dummy variable indicating the gender of student *i* within teacher *j* (female = 1; male = 0). Six dummy variables were created for race/ethnicity, and students who selected white as their race/ethnicity were used as a reference group. ENGLISH is a dummy variable indicating whether or not English is the primary language of student *i* within teacher *j* (English = 1; other language = 0). The modality variables were uncentered, but all other variables were grand–mean centered. The terms u_{0j} and r_{ij} are the error terms associated with the teachers and students, respectively. The results of the conditional HLM of the full data set are shown in Table 5.

Table 5: *Results from the conditional HLM of the entire sample.*

Fixed Effects	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
Intercept, γ_{00}	-0.29	0.05	-6.05	159	<0.001
CBT1, γ_{10}	-0.01	0.03	-0.18	9527	0.86
CBT2, γ_{20}	-0.00	0.03	-0.13	9527	0.90
CBT3, γ_{30}	-0.09	0.03	-3.03	9527	<0.01
Female, γ_{40}	-0.09	0.03	-3.77	9527	<0.001
Hispanic, γ_{50}	-0.34	0.03	-9.75	9527	<0.001
Black, γ_{60}	-0.44	0.04	-11.07	9527	<0.001
Asian, γ_{70}	0.12	0.06	1.79	9527	0.07
American Indian, γ_{80}	-0.28	0.11	-2.63	9527	<0.05
Pacific Islander, γ_{90}	-0.15	0.18	-0.85	9527	0.40
Other, γ_{100}	-0.23	0.04	-6.54	9527	<0.001
English, γ_{110}	0.15	0.05	2.73	9527	<0.05
Random Effects	Standard Deviation	Variance	*d.f.*	χ^2	*p*-value
Intercept, u_0	0.54	0.29	159	2546	<0.001
level-1, *r*	0.98	0.96			

According to the intercept shown in Table 5, the average student performance for students who took the paper version is -0.29 logits (controlling for gender, race/ethnicity, and primary language). The model indicates that there is no significant difference in performance between students who took the paper version and students who took the computer-based version that allowed students to freely navigate through the test (CBT-1) or between students who took the paper version and students who took the computer-based version that restricted test navigation (CBT-2). However, the model does indicate that students who took the computer-based version where they had to click on a “radio” button underneath the item that corresponds to their answer choice selection (CBT-3) scored, on average, 0.09 logits worse than the students who took the paper version (controlling for gender, race/ethnicity, and primary language) ($p < 0.01$).

Grade bands

The same model was run using subsets of the data based on grade band to determine the effect of test modality at the different grade bands. Tables 6, 7, and 8 presents the results of the models from the elementary, middle, and high school students, respectively.

Elementary school students. The results from the model of the data from elementary school students (grades 4 and 5) showed that the average student performance for elementary students who took the paper version is -0.78 logits (controlling for gender, race/ethnicity, and primary language) (see Table 6). The results of this model indicate that there are no statistically significant differences in elementary school students' performance among the testing modalities.

Middle school students. As shown in Table 7, middle school students (grades 6 through 7) scored, on average, -0.37 logits on the paper version (controlling for gender, race/ethnicity, and primary language). As with the elementary school students, no statistically significant differences were found among the different testing modalities for middle school students.

High school students. Table 8 shows the results from the model of the data from the high school students (grades 9 through 12). The average student performance for high school students was 0.00 logits on the paper version (controlling for gender, race/ethnicity, and primary language). This model indicates that high school students scored 0.11 logits worse on the computer-based version where they had to click on a "radio" button underneath the item that corresponds to their answer choice selection (CBT-3) than on the paper version ($p < 0.05$).

Table 6: *Results from the conditional HLM of the elementary school students*.

Fixed Effects	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
Intercept, γ_{00}	-0.78	0.10	-8.10	37	<0.001
CBT1, γ_{10}	-0.05	0.09	-0.59	1306	0.55
CBT2, γ_{20}	-0.04	0.08	-0.51	1306	0.61
CBT3, γ_{30}	-0.09	0.10	-0.99	1306	0.32
Female, γ_{40}	-0.03	0.05	-0.64	1306	0.52
Hispanic, γ_{50}	-0.30	0.07	-4.03	1306	<0.001
Black, γ_{60}	-0.36	0.08	-4.45	1306	<0.001
Asian, γ_{70}	0.03	0.08	0.38	1306	0.70
American Indian, γ_{80}	0.55	0.16	3.53	1306	<0.001
Pacific Islander, γ_{90}	-0.08	0.25	-0.30	1306	0.76
Other, γ_{100}	-0.11	0.08	-1.24	1306	0.21
English, γ_{110}	-0.25	0.10	-2.54	1306	<0.05
Random Effects	Standard Deviation	Variance	*d.f.*	χ^2	*p*-value
Intercept, u_0	0.33	0.11	37	203.30	<0.001
level-1, *r*	0.86	0.74			

Table 7: *Results from the conditional HLM of the middle school students.*

Fixed Effects	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
Intercept, γ_{00}	-0.37	0.04	-8.24	75	<0.001
CBT1, γ_{10}	0.02	0.04	0.43	4094	0.67
CBT2, γ_{20}	0.05	0.04	1.16	4094	0.25
CBT3, γ_{30}	-0.06	0.03	-1.75	4094	0.08
Female, γ_{40}	-0.06	0.04	-1.68	4094	0.09
Hispanic, γ_{50}	-0.27	0.03	-7.99	4094	<0.001
Black, γ_{60}	-0.44	0.05	-8.81	4094	<0.001
Asian, γ_{70}	0.13	0.08	1.57	4094	0.12
American Indian, γ_{80}	-0.41	0.09	-4.53	4094	<0.001
Pacific Islander, γ_{90}	-0.35	0.24	-1.45	4094	0.15
Other, γ_{100}	-0.21	0.05	-4.18	4094	<0.001
English, γ_{110}	0.29	0.08	3.82	4094	<0.001
Random Effects	Standard Deviation	Variance	*d.f.*	χ^2	*p*-value
Intercept, u_0	0.34	0.11	75	595.88	<0.001
level-1, *r*	0.92	0.84			

Table 8: *Results from the conditional HLM of the high school students.*

Fixed Effects	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
Intercept, γ_{00}	0.00	0.08	0.01	79	0.99
CBT1, γ_{10}	0.00	0.05	-0.04	4071	0.97
CBT2, γ_{20}	-0.03	0.05	-0.61	4071	0.54
CBT3, γ_{30}	-0.11	0.05	-2.26	4071	<0.05
Female, γ_{40}	-0.16	0.04	-3.70	4071	<0.001
Hispanic, γ_{50}	-0.41	0.06	-6.30	4071	<0.001
Black, γ_{60}	-0.46	0.07	-6.24	4071	<0.001
Asian, γ_{70}	0.11	0.11	1.05	4071	0.30
American Indian, γ_{80}	-0.72	0.17	-4.17	4071	<0.001
Pacific Islander, γ_{90}	-0.07	0.28	-0.26	4071	0.80
Other, γ_{100}	-0.31	0.06	-4.94	4071	<0.001
English, γ_{110}	0.19	0.07	2.75	4071	<0.05
Random Effects	Standard Deviation	Variance	*d.f.*	χ^2	*p*-value
Intercept, u_0	0.58	0.34	79	1108.80	<0.001
level-1, *r*	1.06	1.13			

Students whose primary language is not English

The following conditional model was run on the data from the students who indicated that English was not their primary language:

$$\text{MEASURE}_{ij} = \gamma_{00} + \gamma_{10}*\text{CBT1}_{ij} + \gamma_{20}*\text{CBT2}_{ij} + \gamma_{30}*\text{CBT3}_{ij} + \gamma_{40}*\text{FEMALE}_{ij} + \gamma_{50}*\text{HI}_{ij} + \gamma_{60}*\text{BL}_{ij} + \gamma_{70}*\text{AS}_{ij} + \gamma_{80}*\text{AI}_{ij} + \gamma_{90}*\text{PI}_{ij} + \gamma_{100}*\text{OT}_{ij} + u_{0j}+ r_{ij}$$

The results of the model, shown in Table 9, indicate that, on average, students whose primary language is not English scored -0.46 logits on the paper version (controlling for gender and race/ethnicity). No statistically significant differences in student performance among the test modalities were found for this subgroup.

Table 9: *Results from the conditional HLM of the students whose primary language is not English.*

Fixed Effects	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
Intercept, γ_{00}	-0.46	0.11	-4.24	119	<0.001
CBT1, γ_{10}	-0.05	0.12	-0.46	560	0.64
CBT2, γ_{20}	-0.03	0.17	-0.18	560	0.86
CBT3, γ_{30}	-0.10	0.13	-0.76	560	0.45
Female, γ_{40}	-0.05	0.08	-0.58	560	0.56
Hispanic, γ_{50}	-0.28	0.15	-1.87	560	0.06
Black, γ_{60}	-0.42	0.24	-1.75	560	0.08
Asian, γ_{70}	0.34	0.20	1.70	560	0.09
American Indian, γ_{80}	0.04	0.27	0.15	560	0.88
Pacific Islander, γ_{90}	-0.08	0.42	-0.18	560	0.85
Other, γ_{100}	-0.13	0.17	-0.75	560	0.46
Random Effects	Standard Deviation	Variance	*d.f.*	χ^2	*p*-value
Intercept, u_0	0.58	0.34	79	1108.80	<0.001
level-1, *r*	1.06	1.13			

Effect Sizes

Table 10 presents the effect sizes for the different test modalities for the entire data set and for the different subgroups. The largest effect sizes were for the comparisons between the PBT and CBT-3 and between CBT-1 and CBT-3 (ranging from 0.04 to 0.10). The smallest effect sizes were for the comparisons between CBT-1 and CBT-2 (ranging from 0.00 to 0.03).

Table 10: *Effect sizes*

	Overall (n = 9698)	Elementary (n = 1355)	Middle (n = 4181)	High (n = 4162)	Non-English (n = 690)
PBT vs. CBT1	0.00	0.06	0.02	0.00	0.05
PBT vs. CBT2	0.00	0.05	0.05	0.03	0.02
PBT vs. CBT3	0.08	0.10	0.06	0.09	0.10
CBT1 vs. CBT2	0.00	0.01	0.03	0.02	0.02
CBT1 vs. CBT3	0.07	0.04	0.08	0.09	0.05

Student preferences

At the end of the computer-based tests, we asked students if they preferred computer-based or paper-based tests. Table 11 summarizes the responses to this question by version of CBT and grade band. Approximately half of the elementary and middle school students preferred computer-based tests over paper-based tests regardless of which computer-based version they took. More high school students preferred paper-based tests over computer-based test. We also found that about a quarter of the students in our study did not have a preference.

Table 11: *Student preferences*

Version	Preference	Elementary	Middle	High
CBT-1	Prefer computer	55%	50%	33%
	Prefer paper	19%	25%	37%
	No preference	26%	25%	30%
CBT-2	Prefer computer	55%	52%	34%
	Prefer paper	24%	26%	40%
	No preference	21%	22%	26%
CBT-3	Prefer computer	56%	45%	35%
	Prefer paper	28%	27%	37%
	No preference	17%	28%	28%

Discussion

Comparing Paper-Based and Computer-Based Tests

The most direct comparison between computer-based test and paper-based test in our study is the comparison between the PBT and CBT-1 (where students could skip and directly click on an answer choice). When comparing the results of these two modalities, we found no statistically significant difference in student performance for the entire sample or for any of the subgroups. Additionally, the effect sizes were either zero or small (i.e., < 0.20; Cohen, 1988). Therefore, although the difference between PBT and CBT-3 was statistically significant for the entire sample and for high school students, because the effect sizes are so small, we conclude that the student scores on the paper-based version and the computer-based version should be considered equivalent.

Comparing Computer-Based Test Versions

Our study was designed so that we could test specific features of our computer-based testing system. A comparison between CBT-1 and CBT-2 provides insight into the effect of being able to skip items and return to previous items. A comparison between CBT-1 and CBT-3 provides insight into the effect of the mode of answer choice selection, that is, whether the students selected an answer choice by directly clicking on the answer choice or by clicking on a corresponding radio button located underneath the item.

Test navigation. Previous comparability studies have found the option to skip, review, and change previous responses had no statistical effect on student performance of college students (Eaves & Smith, 1986; Harvey, 1987; Luecht, Hadadi, Swanson, & Case, 1998). Our finding that there were no significant differences in performance for elementary, middle, or high school students between the CBT-1 and CBT-2 versions is consistent with these studies. The effect sizes for the entire sample and for the different subgroups were either zero or small (i.e., < 0.20; Cohen, 1988).

Answer selection. Research on multiple-choice selection interfaces is sparse, but it is reasonable to think that having to match an answer to a corresponding letter at the bottom of the screen adds an additional level of cognitive processing, which may result in lower student performance. Others have suggested that marking an answer in a different location than where the item appears could be challenging for younger students, students with poor organizational skills, students who have difficulties with concentration, or students who are physically impaired (Dolan et al., 2010). In our study, the effects sizes for the comparisons between PBT and CBT-3, and between CBT-1 and CBT-3 were the largest effect sizes and were significant for the entire sample and for high school students. However, these values (ranging from 0.04 to 0.10) were also small (i.e., < 0.20; Cohen, 1988).

Conclusions

Our findings indicate that, for our instruments, scores on the computer-based and paper-based test versions should be considered equivalent. Regarding the design of computer-based tests, our results suggest that computer-based tests should be designed so that students can click directly on an answer choice to select it. Additionally, our results suggest that restricting test navigation does not affect student performance on computer-based tests and that elementary and middle school students prefer to take computer-based tests over paper-based tests.

One limitation of our study is that our instruments were of low stakes to the students who participated in the study. They were told that their performance on the test would not affect their grade or be shared with the teachers or parents. In a high stakes testing situation, it possible that students may be more likely to want to return to previous items. This could lead to a possible performance difference between computer-based tests that allow skipping and those that restrict test navigation. Also, minor differences in performance on different modalities may be important when the scores are used to make high stakes decisions. So, while our results indicate that these computer-based and paper-based scores can be considered equivalent, we recommend that a comparability study be done for any test that will be administered in multiple modalities.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A120138 to the American Association for the Advancement of Science. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Andrich, D., Marais, I., Humphry, S. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics*, *37*(3), 417-442.

American Association for the Advancement of Science. (n.d.). AAAS Project 2061 Science Assessment Website. Retrieved from www.assessment.aaas.org

Arditi, A., & Cho, J. (2005). Serifs and font legibility. *Vision Research*, 45(23), 2926–2933.

Bernard, M. L., Chaparro, B. S., Mills, M. M., & Halcomb, C. G. (2002). Examining children's reading performance and preference for different computer-displayed text. *Behaviour & Information Technology*, *21*(2), 87–96.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance. *Applied Measurement in Education*, *16*(3), 191–205.

Choi, S. W., & Tinkler, T. (2002, October). Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., New York: Academic.

DeBoer, G. E., Quellmalz, E. S., Davenport, J. L., Timms, M. J., Herrmann-Abell, C. F., Buckley, B. C., Jordan, K.A., Huang, C.W., & Flanagan, J. C. (2014). Comparing three online testing modalities: Using static, active, and interactive online testing modalities to assess middle school students' understanding of fundamental ideas and use of inquiry skills related to ecosystems. *Journal of Research in Science Teaching*, *51*(4), 523–554.

Dolan, R. P., Burling, K. S., Rose, D., Beck, R., Murray, E., Strangman, N., Jude, J., Harms, M., Way, W., Hanna, E., Nichols, A., & Strain-Seymour, E. (2010). Universal Design for Computer-Based Testing (UD-CBT) Guidelines.

Eaves, R. C., & Smith, E. (1986). The Effect of Media and Amount of Microcomputer Experience on Examination Scores. *The Journal of Experimental Education*, *55*(1), 23–26.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2000). The Effect of Computer-Based Tests on Racial/Ethnic, Gender, And Language Groups. *ETS Research Report Series*, *2000*(1), i–17.

Hardcastle, J., Herrmann-Abell, C.F., & DeBoer, G.E. (2017, April-May) Comparing Student Performance on Paper-and-Pencil and Computer-Based-Tests. *Paper presented at the NARST Annual Conference*. San Antonio, TX.

Harvey, A. L. (1987). *Differences in Response Behavior for High and Low Scorers as a Function of Control of Item Presentation on a Computer-Assisted Test.* ETD collection for University of Nebraska - Lincoln. University of Nebraska - Lincoln.

Herrmann-Abell, C.F. & DeBoer, G.E. (2018) Investigating a Learning Progression for Energy Ideas from Upper Elementary Through High School. *Journal for Research in Science Teaching*, 55(1), 68-93.

Hetter, R. D., Segall, D. O., & Bloxom, B. M. (1994). A Comparison of Item Calibration Media in Computerized Adaptive Testing. *Applied Psychological Measurement, 18*(3), 197–204.

Kingston, N.M. (2009). Comparability of Computer- and Paper-Administered Multiple-Choice Tests for K-12 Populations: A Synthesis. *Applied Measurement in Education, 22*, 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.

Linacre, J. M. (2016). Winsteps® Rasch measurement computer program. Beaverton, Oregon. Retrieved from Winsteps.com

Luecht, R. M., Hadadi, A., Swanson, D. B., & Case, S. M. (1998). Testing the Test: A comparative study of a comprehensive basic sciences test using paper-and-pencil and computerized formats. *Academic Medicine*, *73*(10), 51–53.

Mohamad Ali, A. Z., Wahid, R., Samsudin, K., & Zaffwan Idris, M. (2013). Reading on the computer screen: Does font type has effects on Web text readability? *International Education Studies*, 6(3), 26–35.

Paek, P. (2005). Recent Trends in Comparability Studies. *Pearson Educational Measurement*.

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. *Educational and Psychological Measurement, 68*(1), 5-24.

Welch. C., Dunbar, S., Rickels, H. & Chen K. (2014). A comparative evaluation of online and paper & pencil forms for the Iowa Assessments. Iowa Testing Programs, University of Iowa.