Comparing Student Performance on Paper-and-Pencil and Computer-Based-Tests

Joseph Hardcastle, American Association for the Advancement of Science;

Cari F. Herrmann-Abell, American Association for the Advancement of Science;

George E. DeBoer, American Association for the Advancement of Science

Paper presented at the 2017 AERA Annual Meeting

San Antonio, TX

April, 30, 2017

Abstract

Can student performance on computer-based tests (CBT) and paper-and-pencil tests (PPT) be considered equivalent measures of student knowledge? States and school districts are grappling with this question, and although studies addressing this question are growing, additional research is needed. We report on the performance of students who took either a PPT or one of two different CBT containing multiple-choice items assessing science ideas. Propensity score matching was used to create equivalent demographic groups for each testing modality, and Rasch modelling was used to describe student performance. Performance was found to vary across testing modalities by grade band, students' primary language, and the specific CBT system used. These results are discussed in terms of the current literature and the differences between the specific PPT and CBT systems.

Introduction

With the increased availably of computers, many assessments are being administered as computer-based tests (CBT). CBT provides several advantages over paper-and-pencil tests (PPT) 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 PPT-generated scores can be considered equivalent measures of student performance.

Several studies have compared PPT and CBT. Some have found little to no difference between PPT and CBT (Bridgeman, Lennon, & Jackenthal, 2003; Choi & Tinkler, 2002; Hetter, Segall, & Bloxom, 1994). Others have found student performance to be lower on CBT relative to PPT. These differences in student performance have been linked to technological differences such as a CBT requiring scrolling (Bridgeman et al., 2003; Choi & Tinkler, 2002), and to participate characteristics for example the students'

ethnicity, gender, or primary language (Gallagher, Bridgeman, & Cahalan, 2000). For a review of the current literature on this topic see Leeson, 2006 and Paek, 2005. The research on the comparability of CBT and PPT provide some guidance for what to avoid when creating CBT and PPT, but there is still a need to improve our understanding of best practices for design and administrating equivalent CBT and PPT.

In this study we compared student performance on tests administered using three testing systems: PPT and two CBT systems. Tests that assessed 4th through 12th grade students' understanding of energy concepts were administered using each testing system. A comparison of students from each testing group was done using quasi-experimental design in which propensity score matching was used to form demographically equivalent groups for each testing modality. Rasch analysis was used to estimate item- and student-level measures. Students' performances from different testing modalities were compared to evaluate whether PPT and CBT yielded equivalent measures of student knowledge.

Methodology

Assessment Material

The testing material used in this study consisted of 374 distractor-driven, multiple-choice test items, each having four answer choices. Items assessed student understanding of: (1) energy forms and transformations, (2) energy transfer, (3) energy dissipation, and (4) energy conservation (Herrmann-Abeel & DeBoer, 2016). Because we were testing more items than each student could respond to, we used matrix sampling to develop thirty-four different test forms. Linking items were used so that item characteristics could be compared across forms.

Data Collection

Data were collected during two separate test administrations, in the spring of 2015 and the winter of 2015/2016. During both administrations, instructors could choose either PPT or CBT based on availability of computers in their classroom. Instructors were given testing instructions tailored to the testing modality they were using and a list of frequently asked questions.

A total of 34,068 students participated in testing. All students were enrolled in a science class at the time of testing, but not necessarily in a physical science class. Students who answered fewer than six items were excluded from analysis, resulting in a total of 33,422 students.

Testing Modalities

Table 1 compares the features of the different testing modalities. All tests used the same items with identical text, images, and answer choices. Students who took the PPT option were given an 8.5x11 test booklet and a scan able answer sheet. The test booklet was printed in black and white and used a serif font. Serif font has shown to provide good readability for print media (Mohamad Ali, Wahid, Samsudin, & Zaffwan Idris, 2013).

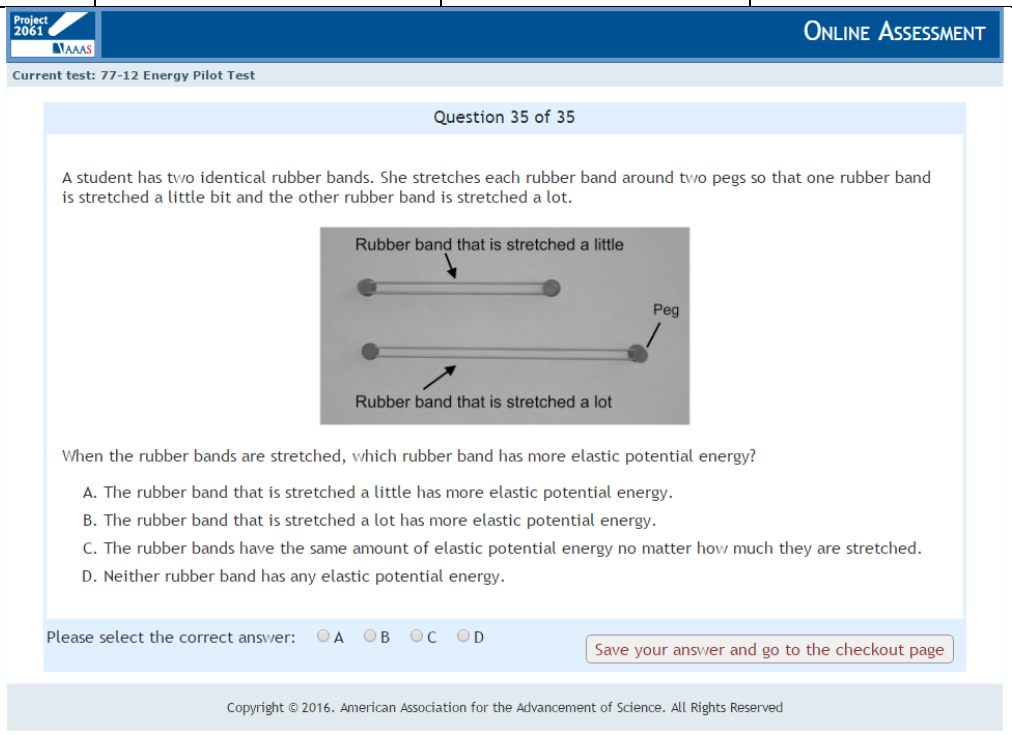
There were two different CBT formats that were used. During the first test administration, the CBT was administered using TAO® (CBT-TAO), an open source online testing system (Open Assessment Technologies, n.d.). During the second administration, the CBT was administered using the AAAS assessment website (CBT-AAAS), where users can create their own tests (American Association for the Advancment of Science, n.d.). Both CBT options included color images and used a sans-serif font. Previous studies have observed no statistical difference in readability when comparing serif and Sans-serif fonts (Arditi & Cho, 2005; Mohamad Ali et al., 2013), however, younger children have indicated a preference for Sans-serif fonts (Bernard, Chaparro, Mills, & Halcomb, 2002).

The main differences between the two CBT options are summarized in Table 1. These include the way the students selected their answer choices, and how students navigated between the items. Screen shots of an item administered using the CBT-TAO and CBT-AAAS testing systems can be seen in Figure 1. Note that in the CBT-AAAS image answers are chosen at bottom of the screen while in the CBT-TAO image answers are chosen by directly clicking the text corresponding to your answer. Another noteworthy y difference between the CBT options was the ability to skip and return to previous items. On the CBT-TAO students could skip items and freely move through the test, while on CBT-AAAS could not allow students to return to previous test items.Table 1

Summary of the differences between each testing modality

	PPT	**CBT-TAO**	**CBT-AAAS**
Font	Serif	Sans-Serif	Sans-Serif
Images	Black and White Images	Some Color Images	Some Color Images
Answer Selection	Students "bubbled" in the letter of their answer on a separate sheet.	Students clicked directly on their answer	Students clicked a "radio" button corresponding to their answer choice.
Order of Items on Test	Fixed	Random for each student	Fixed
Test Navigation	Students could skip items and return to previous	Students could skip items and return to	Students could skip items but could not return to previous

	items.	previous items	item.



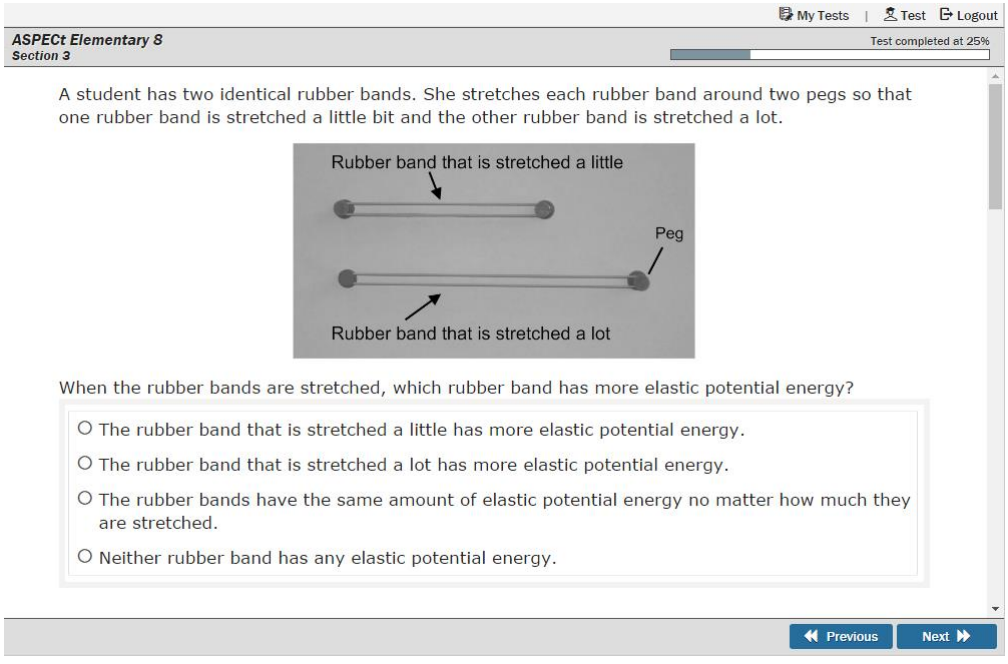


Figure 1.

Images of an example item in the CBT-AAAS (top) and CBT-TAO (bottom).

Propensity Score Matching

Because rather than assigning treatments to classrooms randomly teachers were given the option to administer tests as PPT or CBT, comparable groups were created using propensity score matching (Kim, 2016; Zeng, Yin, & Shedden, 2015). In propensity score matching, individuals in the treatment and control groups are assigned probabilities of being in that group, called their propensity scores, based on various covariates. These probabilities are obtained by fitting the data using the appropriate covariates. Individuals in different groups are then matched according to how similar their probabilities are.

Briefly, matching began by combining data from the two test administrations, consisting of three separate testing groups (PPT, CBT-TAO, and CBT-AAAS). Demographic data (gender, ethnicity, region of the country, and whether English was the student's primary language) were used as covariates to calculate a propensity score for each student in each group, and multi-group matching was used to form equivalent groups (Imbens, 2000; Lechner, 2001). The probabilities for a student being in the PPT, CBT-TAO, and CBT-AAAS groups were obtained using a multi-nominal logistic model. Using the students' probability of being in each group, individual students were matched using "logit nearest neighbor" matching (Wang 2013) with a caliper of 0.2 times the standard deviation of the logit propensity score (Austin 2010). Groups of similar students were matched in a sequential manner, treating the PPT group as a common-reference between the CBT-TAO and CBT-AAAS groups (Rassen et al. 2013). First, equivalent groups of PPT and CBT-AAAS students were formed, during which students that had a distance measure larger than the caliper were removed from the data set. Equivalent groups of PPT and CBT-TAO students were then formed, again removing students that had distance measures larger than the caliper. Calculation of propensity scores and matching was done using custom Python scripts.

Covariate Balance

Three equivalent groups of 4,959 students each were formed using propensity score matching. (Table 2 shows a breakdown of the student demographics by group.) After the matching process, the three groups were compared on a number of relevant covariates. Binary comparisons were made between pairs of groups, and the largest difference was used as an indication of group equivalence. For binary, treatment vs. control group comparisons, a standardized difference can be calculated using the following formula:

$$d = \frac{(p_T - p_C)}{\sqrt{\frac{p_T(1-p_T) + p_C(1-p_C)}{2}}}$$

where p_T and p_C are the prevalence of the trait or feature in the two groups. Results showing the differences before and after propensity matching appear in Table 3. For this work a standardized difference greater than 0.1 was taken to indicate that covariates differed between student populations (Austin, 2009). Prior to matching, 20 covariates had maximum standardized differences larger than 0.1; after matching, only two covariates had standardized differences larger than 0.1.

Table 2
Summary of student demographic variables after propensity score matching

	CBT-TAO (n=4959)	**CBT-AAAS** (n=4959)	**PPT** (n=4959)
Grade Band			
Elementary School	9%	11%	11%
Middle School	48%	50%	48%
High School	43%	39%	41%
Gender			
Male	46%	46%	47%
Female	54%	54%	53%
Primary Language			
English	92%	93%	93%
Other	8%	7%	7%
Race/Ethnicity			
White	55%	54%	56%
Hispanic	16%	16%	14%
Black	10%	10%	7%
Two ethnicities	7%	7%	7%
Three or more ethnicities	6%	5%	7%
Asian	5%	6%	7%
Pacific Island	1%	1%	1%
American Indian	1%	1%	1%
Division of the Country			
Pacific	27%	24%	23%
East North Central	22%	20%	23%
South Atlantic	19%	20%	17%
East South Central	7%	7%	8%
Middle Atlantic	6%	6%	7%
West North Central	6%	5%	5%
West South Central	6%	6%	6%
Mountain	5%	6%	7%
New England	3%	4%	5%

Table 3
Standardized differences before and after propensity score matching

	Standardized Difference	
	Unmatched	Matched
Grade		
4th	0.07	0.035
5th	0.145	0.053
6th	0.129	0.042
7th	0.191	0.034
8th	0.166	0.084
9th	0.153	0.054
10th	0.104	0.049
11th	0.098	0.054
12th	0.076	0.029
Gender		
Male	0.069	0.024
Female	0.069	0.024
Primary Language		
English	0.280	0.008
Other	0.280	0.008
Race/Ethnicity		
White	0.179	0.035
Hispanic	0.185	0.050
Black	0.125	0.113
Two ethnicities	0.161	0.025
Three or more ethnicities	0.067	0.054
Asian	0.040	0.068
Pacific Island	0.024	0.020
American Indian	0.129	0.043
Division of the Country		
Pacific	0.180	0.099
East North Central	0.083	0.083
South Atlantic	0.253	0.069
East South Central	0.276	0.037
Middle Atlantic	0.154	0.031
West North Central	0.314	0.041
West South Central	0.061	0.036
Mountain	0.247	0.062
New England	0.138	0.126

Note: Each standardized difference value is the largest value obtained from pair matching. Values larger than 0.1 are highlighted red.

Rasch Analysis

Rasch analysis was used to estimate student performance and item difficulty using the software package WINSTEPS (Linacre, 2016). In the Rasch model, a student's probability of answering an item correctly is a function of that student's performance and the item's difficulty. Table 4 shows a summary of fit statistics. The lower person separation index (reliability) is due to the fact that there are many fewer data points used to estimate student performance because of the use of matrix sampling. (Each student answered approximately 8% of the total items.)

Table 4
Summary of Rasch Fit Statistics

	Item			Student		
	Min	Max	Median	Min	Max	Median
Standard Error	0.03	0.16	0.07	0.35	1.89	0.40
Infit mean-square	0.85	1.25	0.98	0.46	1.73	0.99
Outfit mean-square	0.71	1.50	0.98	0.28	6.02	0.98
Point-measure correlation coefficients	-0.05	0.53	0.35	-0.74	0.89	0.31
Separation Index (reliability)	9.15 (.99)			1.50 (.69)		

Statistical Analysis

Statistical comparisons between student performance data were made using Python. $P<0.05$ was taken to signify statistically significance. Data were tested for normality using the Shapiro-Wilk test, and the null hypothesis was rejected at the $p<0.05$ level for student performances on the PPT, CBT-AAAS, and CBT-TAO. With the assumption of normality not met for the student performance data, comparisons were made using non-parametric tests. When comparing student performance data from two groups, the Mann-Whitney (MW) test was used, and when comparing data from three groups, the Kruskal-Wallis (KW) test was used.

Results

Comparison of student performance by grade band

Figure 2 shows the average performance of elementary, middle, and high school students for each modality. Elementary and middle school students scored lower on the CBT-AAAS than on the PPT. Elementary students scored .2 logits lower (MW U=128,944, p<.001), and middle school students scored .1 logit lower (MW U=2,706,071, p<.001). Scores on the CBT-TAO were not significantly different from scores on the PPT for middle school students (MW U=2,812,762, p>.05) and were marginally different for elementary school students (MW U=133,853, p<.05). High school students performed similarly on all test formats (KW H=1.51, p>.05).

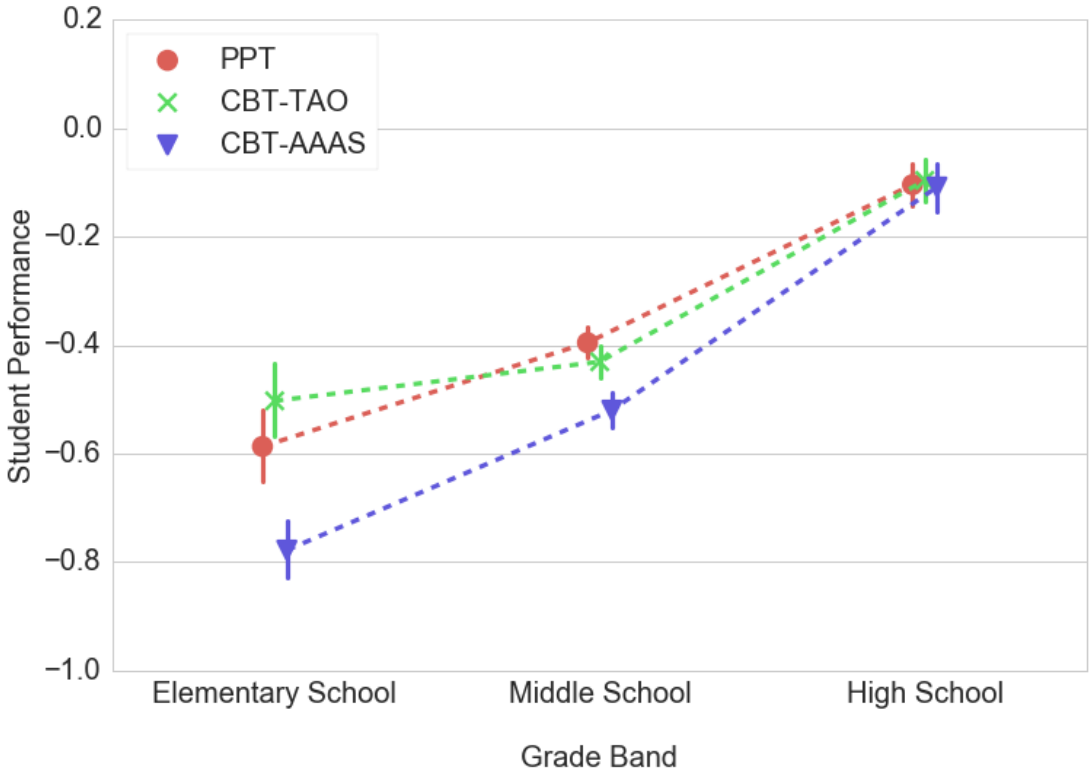


Figure 2.

Average student performance (in logits) of elementary, middle, and high school students who took tests using each testing modality. (Error bars are standard error confidence intervals calculated using bootstrapping of 10,000 bootstrap samples.)

Comparison of student performance by gender

Figure 3 shows the average performance of elementary, middle, and high school male and female students for each testing modality. In both elementary and middle school, male and female students performed lower on the CBT-AAAS compared to the PPT or CBT-TAO. In middle school, males performed slightly higher on the CBT-AAAS compared to their female counterparts (difference in average performance = 0.08 logits, MW U=799,923, p<.05). In high school, the performance of both male and female high school students did not vary with testing modality (KW H=0.66, p>.05 and KW H=1.54, p>.05, respectively).

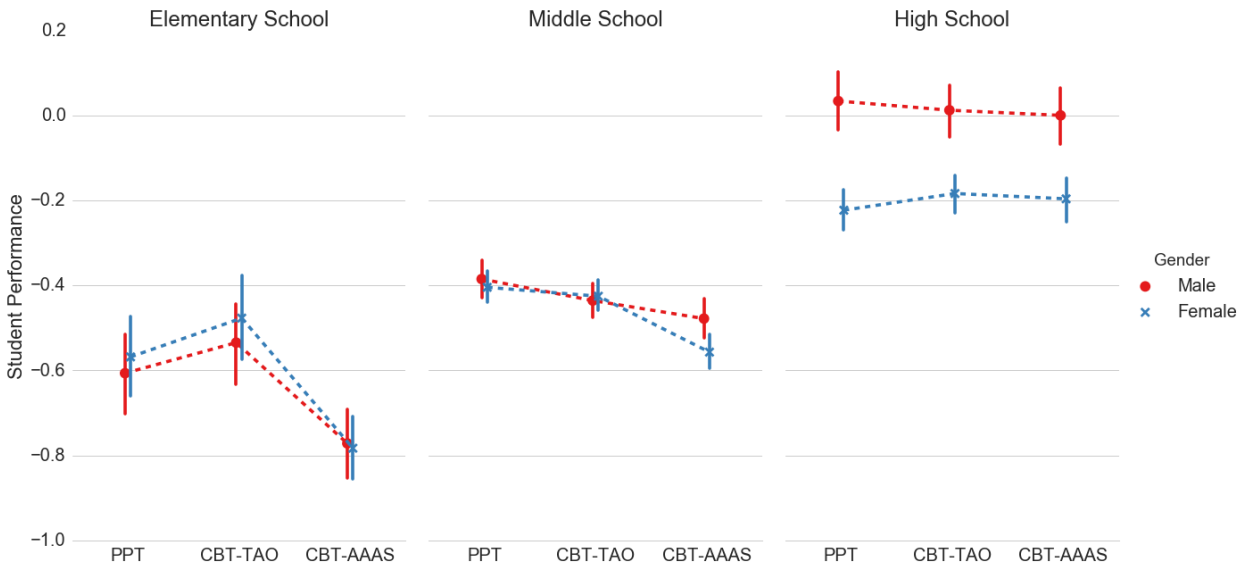


Figure 3.

Average student performance (in logits) of elementary, middle, and high school male and female students who took tests using each testing modality. (Note: Error bars are standard error confidence intervals calculated using bootstrapping of 10,000 bootstrap samples.)

Comparison of student performance by primary language

Figure 4 shows the average performance of students whose primary language is English compared to those who indicated their primary language was not English. Students who indicated English was their primary language performed approximately 0.1 logits lower on the CBT-AAAS compared to the PPT (MW U=9,997,733, p<.001). Students who indicated English was not their primary language performed approximately 0.15 logits lower on both of the CBT-TAO and CBT-AAAS formats compared to the PPT (U=60,941, p<.01 and U=56,971, p<.001, respectively).

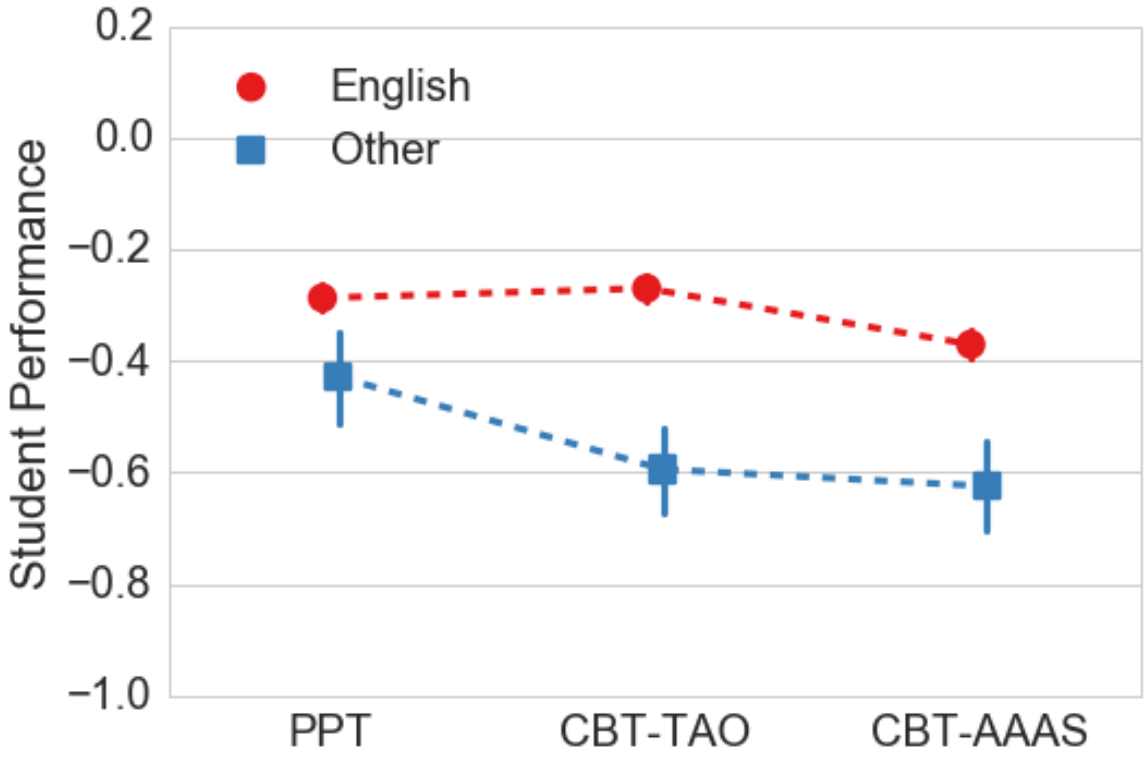


Figure 4.

Average student performance (in logits) of students whose primary language was English and students who indicated their primarily language was not English (Other). (Note: Error bars are standard error confidence intervals calculated using bootstrapping of 10,000 bootstrap samples.)

Discussion

Our results indicate that elementary and middle school students did not perform as well on CBT-AAAS compared to the other testing modalities. A major difference between the two computer tests, CBT-TAO and CBT-AAAS, was that CBT-AAAS did not allow students to return to previous items on the test (Table 1). 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 high school students had equivalent performances on all testing modalities is consistent with these studies; however, our results for elementary and middle school students are not. This may indicate that being able to skip, review, and change previous responses could be beneficial for younger students in elementary and middle school, but have no influence on older students in high school and college.

Another presentation difference between the two CBT modalities was that students clicked directly on their answer choice in the CBT-TAO, whereas students who took the CBT-AAAS were required to choose the letter at the bottom of the screen that corresponded to their answer choice (Table 1 & Figure 1). Research on multiple-choice selection interfaces is sparse, but marking an answer in a different location on a multiple-choice test could be challenging for younger students, students with poor organizational skills, difficulties with concentration, or students who are physically impaired (Dolan et al., 2010). In addition, having to match your answer to a corresponding letter at the bottom of the screen likely adds an additional level of complexity and cognitive processing, which may explain elementary school and middle school students' lower performance on the CBT-AAAS. It is worth noting that the performance gap between CBT-AAAS and the rest of the testing modalities diminished from elementary to middle to become statistically equivalent in high school. We can only speculate as to why no modality effects were observed for high school students, but high school students may be familiar enough with testing and online environments such that particular online testing systems give them neither an advantage nor disadvantage.

Gender was found to have little influence on a student's performance on PPT or CBT; however students whose primary language was not English had lower performances on both CBTs relative to the PPT. Our finding that gender does not play a significant role in student performance on CBT and PPT agrees with previous comparability studies (Karkee, Kim, Fatica, & Mcgraw-hill, 2010; Poggio, Glasnapp, Yang, & Poggio, 2005). The cause of language modality effects are unclear, but could be due to linguistic challenges that the online environment might present or fewer opportunities to use computers in non-English speaking environments.

Our study has several limitations worth noting. Methodologically, quasi-experimental designs like this have limitations when trying to form comparable groups, especially when contrasted with random assignment or a single group design. (A discussion of these limitations can be found in Winter, 2010). Our design is also limited by the CBT systems that were available. Ideally, we would alter a single feature of a CBT system when comparing multiple CBT systems in order to determine how specific features influence student performance. There are also several variables that we were unable to probe, such as eligibility for free or reduced lunch and time on task during testing, that may of interest when comparing PPT and CBT.

Our findings indicate that for high school students PPT and CBT may be considered equivalent measurements of student performance; however, elementary and middle school equivalence depends on the features of the specific computer testing system used. Another place where significant differences were found was for student whose primary language was not English. While we were unable to directly infer which CBT features are important for PPT/CBT equivalence in this study, comparing our results to previous research findings suggests not allowing students to return to previous items on CBT and requiring students to choose their answer choice in a different location on the screen could result in lower performance on a CBT for younger students. These features had no influence on the performance of high school students, suggesting that the equivalence of PPT and CBT may also depend on a student's familiarity with computers or different computer environments.

A more controlled study in which individual testing parameters were varied has been done. Analysis of those results should provide additional information on what features are important for designing computer-based tests equivalent paper-based tests.

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 the views of the Institute or the U.S. Department of Education.

References

American Association for the Advancment 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. http://doi.org/10.1016/j.visres.2005.06.013

Austin, P. C. (2009). Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and monte carlo simulations. *Biometrical Journal*, *51*(1), 171–184. http://doi.org/10.1002/bimj.200810488

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. http://doi.org/10.1080/01449290210146737

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. http://doi.org/10.1207/S15324818AME1603_2

Choi, S. W., & Tinkler, T. (2002). 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*, (October). Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Evaluating+comparability+of+paper-and-pencil+and+computer-based+assessment+in+a+K-12+setting#0

DeBoer, G. E., Quellmalz, E. S., Davenport, J. L., Timms, M. J., Herrmann-Abell, C. F., Buckley, B. C., … 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. http://doi.org/10.1002/tea.21145

Dolan, R. P., Burling, K. S., Rose, D., Beck, R., Murray, E., Strangman, N., … Strain-Seymour, E. (2010). Universal Design for Computer-Based Testing (UD-CBT) Guidelines. Retrieved from http://images.pearsonassessments.com/images/tmrs/tmrs_rg/TMRS_RR_UDCBTGuidelinesrevB.pdf

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. http://doi.org/10.1080/00220973.1986.10806430

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. http://doi.org/10.1002/j.2333-8504.2000.tb01831.x

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. Retrieved from http://digitalcommons.unl.edu/dissertations/AAI8717253

Herrmann-Abeel, C. F., & DeBoer, G. E. (2016). Using Rasch Modeling and Option Probability Curves to Diagnose Students' Misconceptions. *Paper Presented at the 2016 AERA Annual Meeting Washington, DC*.

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. http://doi.org/10.1177/014662169401800301

Imbens, G. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710. http://doi.org/10.1093/biomet/87.3.706

Karkee, T., Kim, D., Fatica, K., & Mcgraw-hill, C. T. B. (2010). Comparability Study of Online and Paper and Pencil Tests Using Modified Internally and Externally Matched Criteria, 1–16.

Kim, J. (2016). Recent Trends of Mode Comparability Studies. In *National Council on Measurement in Education*.

Lechner, M. (2001). Identification and Estimation of Causal Effects Independence Assumption Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumption. *Econometric Evaluation of Labour Market Policies*, (13), 43–58. http://doi.org/10.2139/ssrn.177089

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. http://doi.org/10.1207/s15327574ijt0601_1

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 Bas... : Academic Medicine. *Journal of the Association of American Medical Colleges*, *73*(10), S51–53. Retrieved from http://journals.lww.com/academicmedicine/citation/1998/10000/testing_the_test__a_comparative_study_of_a.43.aspx

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. http://doi.org/10.5539/ies.v6n3p26

Open Assessment Technologies. (n.d.). TAO® Open Source assessment Platform. Retrieved from www.taotesting.com

Paek, P. (2005). Recent Trends in Comparability Studies. *Pearson Educational Measurement. Retrieved*, (August), 1–34. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/5FC04F5A-E79D-45FE-8484-07AACAE2DA75/0/TrendsCompStudies_rr0505.pdf

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). Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=EJ848516\nhttp://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ848516

Winter, P. C. (2010). Evaluating the comparability of scores from achievement test variations.

Zeng, J., Yin, P., & Shedden, K. A. (2015). Does Matching Quality Matter in Mode Comparison Studies? *Educational and Psychological Measurement*, *75*(6), 1045–1062. http://doi.org/10.1177/0013164414565006