Comparability of Paper and Computer Administrations in Terms of Proficiency Interpretations

A paper presented at the annual meeting of the National Council on Measurement in Education
New Orleans, LA

Shalini Kapoor

Catherine Welch

April 2011

Abstract

This study compares students' performance on paper and pencil (PPT) and computer-based test (CBT) on a large-scale statewide Mathematics assessment and discusses the impact of mode of administration on proficiency category classifications of students. Analyses conducted at grade levels five and eight indicate average grade five students found the PPT slightly easier than CBT and grade eight students' found the CBT slightly easier than PPT. Classification consistency results suggest that the mode of administration did not affect students' classification in the proficiency categories.

Keywords: Computer-based testing; Proficiency cut points; Computer vs. Paper administration; online assessment; Decision Consistency at cut points.

Background and Purpose

Advantages of computer based tests (CBT) such as quick turnaround of results, reduction in mailing and paper costs, and increase in student motivation, have encouraged the transition from paper-pencil tests (PPT) to CBTs. The quick turnaround of results gives teachers timely and valuable information to tailor their instruction for the remainder of the school year (Peak, 2005; Bennett,2003). In 2006-2007, while 23 states were reported to offer computer tests as part of their K-12 large-scale assessments, many others were conducting pilot tests to decide whether to make the transition (Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008). The high stakes nature of these assessments, prompted states to do comparability studies of the two modes: PPT and CBT, to ensure that neither mode advantages or disadvantages particular students. In accordance with AERA, APA, & NCME (2004) Standard 4.10, before a new mode is used, empirical evidence is required to make sure it does not disadvantage any students.

Comparability across modes of administration can be gauged by examining item, test, construct, and/or skill score characteristics. Studies across grade levels, content areas, types of tests and items, general administration, presentation characteristics, and response requirements, have compared modes of administration (Peak, 2005; Bennett, 2003). K-12 comparability studies (Peak, 2005) in general, show little or no effect of mode of administration on performance across grades and academic subjects. Two areas where differences remain are items relating to long passages in reading and graphical questions in mathematics (Peak, 2005). A meta-analysis by Wang, Jiao, Young, Brooks, & Olson (2007) found no significant difference in performance in the two modes in mathematics. In contrast, Choi and Tinkler (2002) and Bennett (2001) found that CBT items were generally harder than PPT items across subject areas.

Mode effects for PPT and CBT should be empirically examined, and documented. Depending upon the use of the results, these mode effects may affect the scale conversions and reporting metrics. Mode effects could be due to testing conditions, test scoring, test questions or examinee groups (Kolen, 1999). The majority of comparability studies have focused on differences in means and standard deviations of test scores with little focus on precision issues (Peak, 2005). If two modes are comparable, they should have the same measurement precision across ability/proficiency levels. Comparing score distributions alone may be misleading (Lottridge, Nicewander, Schulz, & Mitzel, 2008). Even though the overall score distributions produced by the modes may be equal, it is possible an examinee's score differs substantially between the two modes. This could change the proficiency category of an individual student. Mode of administration could confound the percent of students at or above a certain proficiency / achievement level (Lottridge, et al, 2008).

This paper focuses on the comparability of CBT and PPT administration in terms of proficiency/achievement level interpretations and classification consistency at proficiency cut scores. Specifically the paper will address the question – what is the impact on proficiency classifications when a state assessment moves from one mode (PPT) to another (CBT) or to support both modes simultaneously? The three proficiency levels are basic, proficient, and advanced, and the proportion of students classified differently due to mode effects in each proficiency category are examined.

Method

Sample

A CBT pilot study was conducted in spring 2010 in a Midwestern state. Schools from sixty-one school districts in the state voluntarily participated in the pilot study. All schools took the required PPT mathematics assessment in the academic year 2009-10 and were invited to participate in the CBT pilot. The pilot study was conducted at two grade levels: five and eight. The sample in this analysis consisted of 689 fifth grade students (52.5% males and 47.5% females), and 676 eighth grade students (48.1% males and 51.9% females).

Instrument

Two tests were assembled from the same pool of field test items to match the same content and technical specifications. One test was administered as a PPT and the second as a CBT. The two tests consist of items in the following content areas: math concepts, estimation, problem solving, and data interpretation. Table 1 gives the summary statistics of the two tests at the different grade levels in raw score percentages. For the PPT, there were 66 items in grade five and 81 in grade eight. In the CBT, there were 60 items at both grade levels.

Training tools, tutorials, and practice experiences supported the CBT administration delivered via the internet. Students were provided online tools such as eraser, highlighter, ruler, review button, summary options, striker, and a pause button. Review buttons were accessible for marking an item if they wanted to come back to that item. Strikers were available for crossing out answer choices that students thought were wrong. At the end of the test, the summary button could give details of the items attempted, skipped, or marked for review. Navigational tools such as review and summary buttons simulate test taking strategies students use when taking PPT, this results in more equivalent scores on CBTs (Peak, 2005). The PPT was administered under standardized conditions. Students could review answers and had to bubble answers in an answer booklet. Calculators were permitted for the PPT.

Procedure

This study has a single group design as all students took both CBT and PPT. Schools took the PPT as part of the statewide accountability assessment in the spring of 2010 and the CBT pilot administration was conducted within two weeks following the statewide assessment.

Classical and Item Response Theory (IRT) statistics were generated to compare the results on the two assessments. Total test performance in terms of mean p-values, standard deviations, and range for the two assessments were compared. Reliability estimates (KR20) for each of the two tests were calculated and the correlation and disattenuated correlation between performances on the two tests were estimated. Item parameters were estimated using IRT three-parameter logistic (3PL) model to compare the test and item characteristics of the two tests.

PPT and CBT were linked to establish proficiency cut scores on the CBT comparable to the cut scores on the PPT. Linking was done using the equipercentile method with cubic spline smoothing of the equating function and then, to get the cut score on the CBT, the raw score equivalent on the CBT linked scores corresponding to the PPT cut score was estimated. These cut scores were then used to gauge mode effect on the proportion of students in the proficiency categories. Finally, classification consistency indices were estimated for the three proficiency categories: basic, proficient, and advanced.

Classification consistency (CC) refers to the extent to which examinees are classified into same categories over replications of similar measurement procedures (Lee, 2010). CC indices are based on a comparison of scores for two observed score distributions. However, if scores from only a single administration are available then CC indices are calculated by estimating the observed score distribution by imposing a psychometric model on the test data (Lee, 2010). CC indices were estimated for single and double administration using binary and multiple classification. Binary classification is when each examinee is classified into one of the two categories i.e. CC indices calculated separately for basic/proficient and proficient/advanced and multiple classification is when each examinee is classified into one of the three or more categories / achievement levels i.e. basic, proficient and advanced in this study.

Single administration CC indices i.e. CC indices for PPT if it was the only test administered and CC indices for CBT if it was the only test administered, were calculated to compare them to those calculated from double administration (when both CBT and PPT were administered). In this study, single administration CC indices were estimated using the beta binomial model. This model assumes the forms are randomly parallel. CBT and PPT constructed from the same item pool are also consistent with the randomly parallel assumption. Hence, any differences in the CC indices for single and double administration could be attributed to mode effects. BB-CLASS computer program was used to estimate single administration CC indices by the Hansen and Brennan (HB) procedure. This procedure uses the observed score distributions predicted from a psychometric model to estimate classification consistency indices and assumes that a test consists of equally weighted, dichotomously-scored items (Brennan, BB-CLASS, 2004). In this study, the true score distribution was modeled on the four-parameter beta distribution and the conditional error distribution on the binomial distribution.

Estimating CC indices for the double administration involves counting the proportion of examinees assigned to each classification category on both tests (Lee, 2010). Double administration classification indices for CBT and PPT are given Tables 3a and 3b.

Results

CBT and PPT (Item and Test Characteristics)

Average proportion correct (p-values) and score distributions (means and standard deviations) were compared across modes. Table 1 indicates similar reliability coefficients (KR20), score distributions (raw score percent metric), and average p-values for CBT and PPT for both grade levels.

Table 1: Summary statistics of CBT and PPT

		Grade 5		Grade 8	
		PPT	**CBT**	**PPT**	**CBT**
N		689	689	676	676
Raw Score Percent Correct	Mean	67.01	65.62	68.55	70.18
	SD	16.85	15.45	17.04	17.52
	Min	21.21	21.67	18.52	15
	Max	100	98.33	100	100
Average p-value		0.67	0.66	0.69	0.70
Reliability (KR-20)		0.91	0.89	0.93	0.92
Correlation (PPT & CBT)		0.86		0.90	
Disattentuated Correlation (PPT & CBT)		0.96		0.97	

In Table 1, a higher mean score for percent correct on PPT compared to CBT in grade five indicates students found PPT to be slightly easier than CBT. In grade eight, students found CBT to be slightly easier than PPT. After correcting for attenuation the correlation between CBT and PPT for grade five was 0.96 and for grade eight was 0.97. The high values of these disattenuated correlations suggest that the two tests are measuring the same thing. Average p-values (proportion correct) were similar across the two modes at the two grade levels.

Item parameters were estimated to compare item characteristics across modes. Table 2 gives the average parameter estimates (â, $\hat{b}$, and ĉ) of CBT and PPT at the two grade levels. These were calculated using the 3PL IRT model. A 3PL model was used as guessing is expected in a multiple choice mathematics test. Item parameters of the two tests were calibrated separately using the statistical package "irtoys" in R software (version 2.10.1). This package provides an interface for estimation of item parameters using different programs like ICL and BILOG. ICL was used in this study.

Table 2: CBT and PPT average parameter estimates

		Grade 5		**Grade 8**	
		PPT	**CBT**	**PPT**	**CBT**
Average	â	0.81	0.85	1.45	1.57
	$\hat{b}$	-0.52	-0.39	-0.40	-0.50
	ĉ	0.21	0.21	0.20	0.17

Similar item parameter estimates (Table 2) and summary statistics (Table 1) indicate the two modes are comparable in terms of overall results. These statistics, similar score distributions (means and standard deviations) and the two tests being constructed from the same item pool to the same technical and content specifications suggest these tests are measuring the same construct (mathematics, in this case). High KR20s, 0.91 (PPT) and 0.89 (CBT) at grade five level and 0.93 (PPT) and 0.92 (CBT) at grade eight level indicate the tests are internally consistent.

Proficiency Interpretations

To examine the number of students classified differently at the two cut scores for each grade, the PPT and the corresponding CBT raw scores for each student were graphed (see figures 3a and 3b). On the x-axis are the PPT raw scores and on the y-axis the CBT raw scores.

The gray and black points in Figures 3a and 3b represent students' raw scores on CBT and PPT. The vertical lines represent the cut scores (in the raw-score metric) on the PPT corresponding to the percentile ranks used for the cut points at the three proficiency categories. The dotted horizontal line represents the CBT cut score (in the raw-score metric) equivalent to the PPT cut score computed by linking the two tests. The black solid circles represent students classified differently due to the effect of mode of administration and errors of measurement at the basic/proficient cut score. The black solid triangles represent students classified differently at the proficient/advanced cut score. Table 3a and 3b give the percentage of students classified differently at the two cut scores at each grade level.

Figure 3a

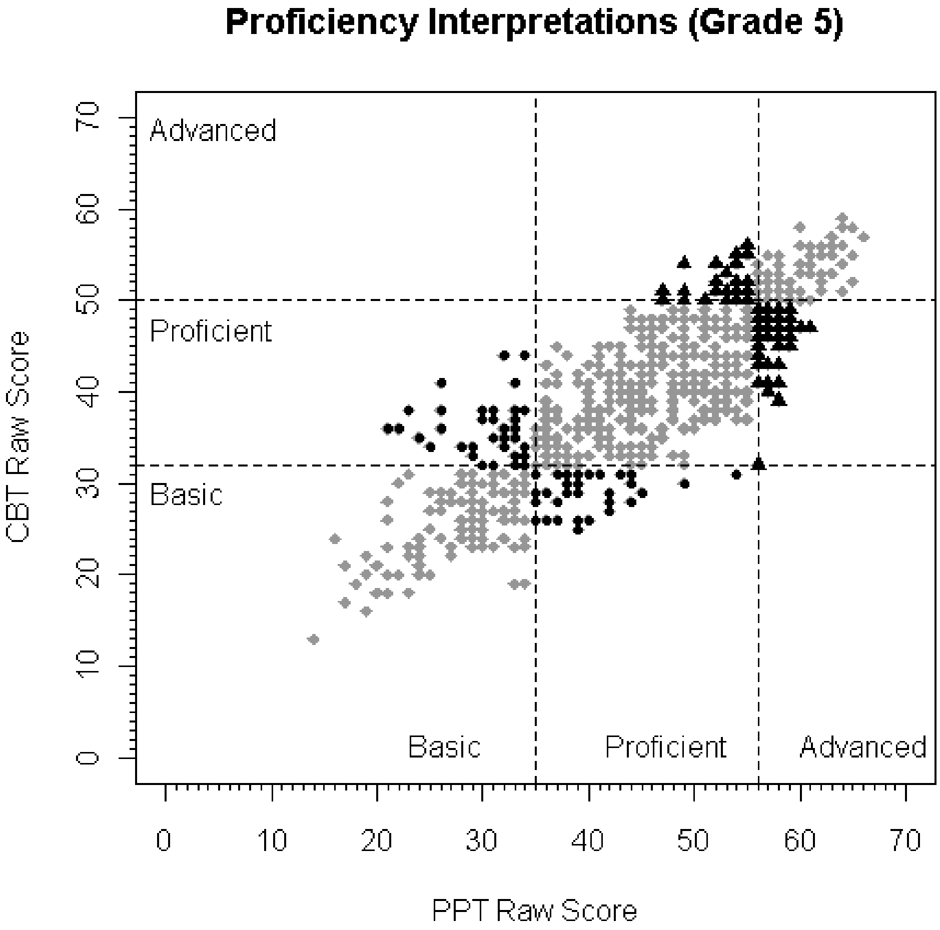

Table 3a: Proportion of consistent / inconsistent classifications

Grade 5 (N=689)		**PPT**		
		Basic	**Proficient**	**Advanced**
CBT	**Advanced**	-	**0.0435**	**0.1205**
	Proficient	**0.0595**	**0.4964**	**0.0610**
	Basic	**0.1626**	**0.0566**	-

Proportion of students classified in the same category = 0.1626 + 0.4964 + 0.1205 = **0.7795**

Proportion of students classified differently = 0.0595 + 0.0566 + 0.0435 + 0.0610 = **0.2206**

Figure 3b

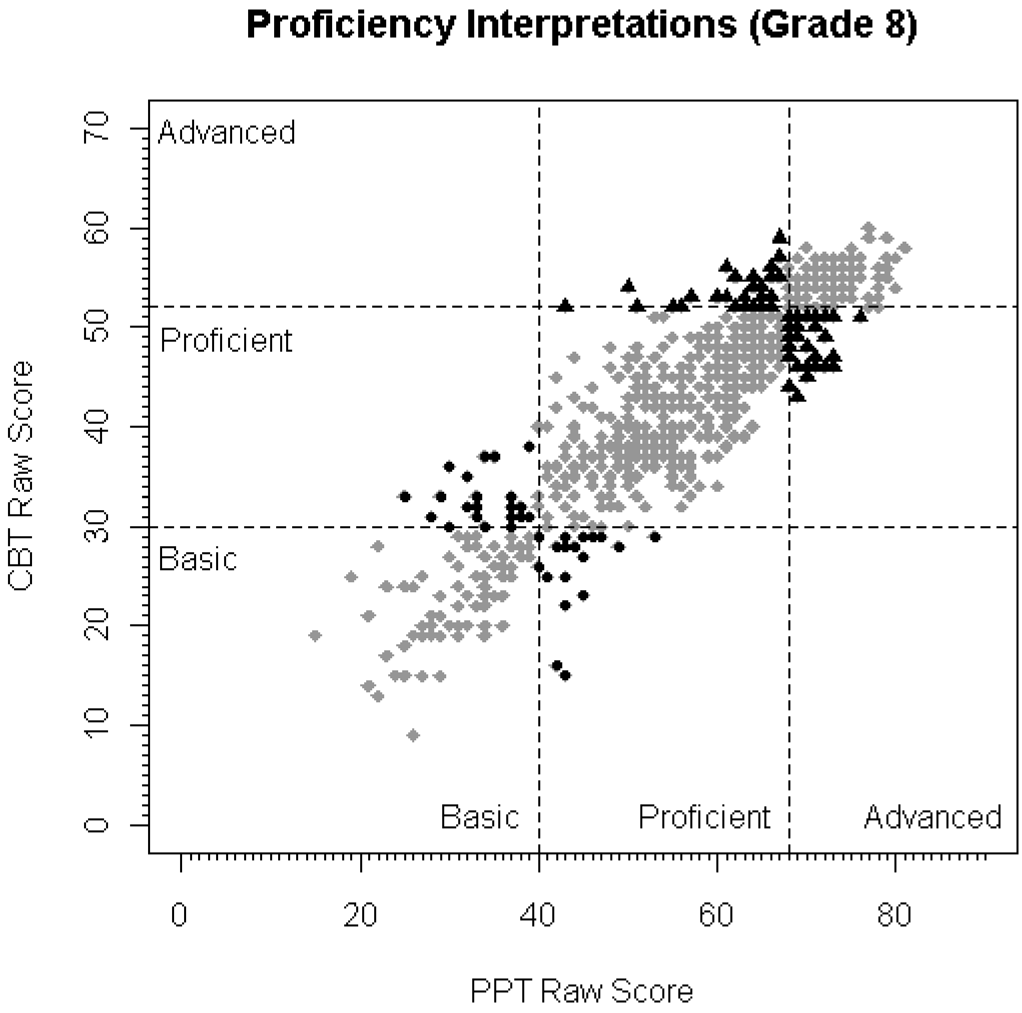

Table 3b: Proportion of consistent / inconsistent classifications

Grade 8 (N=676)		**PPT**		
		Basic	**Proficient**	**Advanced**
CBT	**Advanced**	-	**0.0562**	**0.1760**
	Proficient	**0.0340**	**0.5414**	**0.0503**
	Basic	**0.1109**	**0.0311**	-

Proportion of students classified in the same category = 0.1109 + 0.5414 + 0.1760 = **0.8284**

Proportion of students classified differently = 0.0340 + 0.0562 + 0.0311 + 0.0503 = **0.1716**

Table 3a indicates approximately 5.66% of grade five students were classified as proficient on the PPT but as basic on CBT. While 5.95% classified as proficient on CBT moved to the basic category on PPT. In addition, 6.10% students moved up to the advanced category on the PPT however, the same students were in the proficient category on the CBT. Table 3b indicates similar movements in grade eight. At the eighth grade, a total of 17.16% of the students were classified differently on the two tests. 6.51% (3.11% + 3.40%) at the basic / proficient cut score and 10.65% (5.62% + 5.03%) at the proficient /advanced cut score. To investigate further, classification consistency indices were calculated for single administration (PPT and CBT separately) and their values were compared to double administration (CBT and PPT together).

Classification Consistency (CC)

High CC indices indicate a high agreement in students' classification in proficiency categories over replication of similar measurement procedures. CC was generally high across grades and tests as indicated in Tables 4 and 5. For single administration (binary classifications: Table 4) at the basic/proficient level, CC is higher for grade eight and at the proficient/advanced level, it is higher for grade five. For multiple classifications (Table 5), CC is higher for grade eight than grade five. Proportion of CC under multiple classifications is lower than binary classification at both grade levels.

Table 4 (Binary Classification)

Proportion of Consistent classification (Binary classification)

Grade 5	Single administration		Double administration
	PPT	CBT	PPT & CBT
Basic / Proficient	0.8980	0.8731	0.8839
Proficient / Advanced	0.9000	0.8981	0.8955

Proportion of Consistent classification (Binary classification)

Grade 8	Single administration		Double administration
	PPT	CBT	PPT & CBT
Basic / Proficient	0.9289	0.9264	0.9349
Proficient / Advanced	0.8999	0.8813	0.8935

Table 5 (Multiple Classification)

Proportion of Consistent classification (Multiple classifications)

Grade 5	Single administration		Double administration
	PPT	CBT	PPT & CBT
Basic / Proficient & Proficient / Advanced	0.7980	0.7712	0.7795

Proportion of Consistent classification (Multiple classifications)

Grade 8	Single administration		Double administration
	PPT	CBT	PPT & CBT
Basic / Proficient & Proficient / Advanced	0.8288	0.8077	0.8284

Table 4 and 5 show no consistent patterns across single and double administration in both grades. This indicates absence of a mode effect on classification of students in proficiency categories in the two tests under consideration here. The students classified differently indicated by the black circles and triangles in Figure 3a and 3b could be due to errors of measurement.

Summary and Discussion

Differences in CBT and PPT could be attributed to a number of sources such as mode of administration, examinees' exposure to computers at school and home, computer software familiarity, and administration conditions. In this study, the focus was on mode of administration. Future research could examine other sources of variations in PPT and CBT. Additional possible future research could replicate this study for other content areas, grade levels, use other models such as two-parameter beta binomial and IRT based models for estimating classification consistency indices, and replicating the study in the scale score metric.

This study was conducted in a Midwestern state in the mathematics content area. This limits the generalizability of the study results. According to Kolen (1999), "tests are administered to examinee groups". It is possible that scores on two tests could be comparable for one group but not for another. These examinee groups can differ in demographic characteristics, computer exposure, motivation, and educational characteristics (Kolen,1999). Single group design used in this study, is the richest source of information in comparability studies (Lottridge, et al, 2008). However, limitation of this design is taking two tests may not generalize to a population where only one test is administered and problems due to fatigue and motivation caused by testing the same students twice may occur.

Given the increasing numbers and usage of computers in schools and at homes, it appears that student performance can be accessed using technology and computers (Peak, 2005). However, making sure students are not disadvantaged in any way; especially at the proficiency cut scores where high stakes decisions are made has to be ensured. Incomparable scores across delivery modes can lead to incorrect decisions. For example, under No Child Left Behind, schools could under or overestimate their standing with respect to Adequate Yearly Progress (Bennett, 2003).

In most schools, decisions regarding enrollment in advanced or remedial classes are based, partially, on results from large-scale assessments. If mode effects interact with cut scores students can be inappropriately advantaged or disadvantaged. This raises questions regarding equity if states allowed schools/building to select multiple modes or opt for one over the other. This study encourages test developers, when conducting comparability studies of PPT and CBT or other modes of delivery, to include analyses at proficiency level cut scores.

References

Bennett, R. E. (2003). *Online Assessment and the Comparability of Score Meaning*. Princeton: ETS.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my Mathematics Test on Computer? A second Empirical Study of Mode Effects in NAEP. *The Journal of Technology, Learning and Assessment*, 6(9).

Brennan, R. L. (2004, December). Manual for BB-CLASS. *CASMA Computer Programs*. University of Iowa.

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

IOWA Department of Education. (n.d.). *NCLB, Accountability Workbook & Application*. Retrieved July 20, 2010, from www.iowa.gov: http://www.iowa.gov/educate

Kolen, M. K. (1999). Threats to Score Comparability With Applications to Performance Assessments and Computerized Adaptive Tests. *Educational Assessment,* 73-96.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking.* New York: Springer.

Lee, W.C. (2010). Classification Consistency and Accuracy for Complex Assessments Using Item Response Theory. *Journal of Educational Measurement*, 1-17.

Lottridge, S., Nicewander, A., Schulz, M., & Mitzel, H. (2008). *Comparability of Paper-based and Computer-based Tests: A Review of the Methodology.* Monterey: Pacific Metrics Corporation.

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

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*, 219-238.

Yi, H. S., Kim, S., & Brennan, R. L. (2007). A Method for Estimating Classification Consistency Indices for Two Equated Forms. *Applied Psychological Measurement*, 275-291.