J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 6, Number 4 · December 2007

Comparability of Computer and Paper-and-Pencil Versions of Algebra and Biology Assessments

Do-Hong Kim & Huynh Huynh

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 6, Number 4

Comparability of Computer and Paper-and-Pencil Versions of Algebra and Biology Assessments

Do-Hong Kim & Huynh Huynh

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Jennifer Higgins
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Kim, D.-H., & Huynh, H. (2007). Comparability of Computer and Paper-and-Pencil Versions of Algebra and Biology Assessments. *Journal of Technology, Learning, and Assessment, 6*(4). Retrieved [date] from http://www.jtla.org.

J·T·L·A

Abstract:

This study examined comparability of student scores obtained from computerized and paper-and-pencil formats of the large-scale statewide end-of-course (EOC) examinations in the two subject areas of Algebra and Biology. Evidence in support of comparability of computerized and paper-based tests was sought by examining scale scores, item parameter estimates, test characteristic curves, test information functions, Rasch ability estimates at the content domain level, and the equivalence of the construct. Overall, the results support the comparability of computerized and paper-based tests at the item-level, subtest-level, and whole test-level in both subject areas. No evidence was found to suggest that the administration mode changed the construct being measured.

J·T·L·A

Comparability of Computer and Paper-and-Pencil Versions of Algebra and Biology Assessments

Do-Hong Kim
University of North Carolina at Charlotte
Huynh Huynh
University of South Carolina

Introduction

The history of computerized testing began in the early 1970s (Drasgow, 2002). Limited computer capability and high costs, however, have limited the implementation of computerized testing. With the advent of new technologies, computerized testing has begun to be developed and implemented in large-scale testing programs such as licensure, certification, admissions, and psychological tests. For example, the Graduate Record Examinations (GRE) has been administered in computer-adaptive format for several years. In 1998, the Test of English as a Foreign Language (TOEFL) began transitioning to computer-adaptive testing. Recently, the new TOEFL iBT began administration via the Internet in a non-adaptive format. Increased testing requirements and tight deadlines imposed by the No Child Left Behind Act of 2001 (NCLB) (Public Law No: 107–110) have led to new ways that states can measure student performance more efficiently. Given several benefits that computer-based testing (CBT) can offer over traditional paper-and-pencil testing (PPT), CBT has been a popular choice for statewide assessment programs in order to meet these increased demands for school accountability. For example, 21 states and the District of Columbia offered computerized testing in the 2005–06 school year (Swanson, 2006). Some states have employed or piloted computer-based assessments for the purpose of improving instruction, monitoring student progress, and promoting accountability, while other states have employed it as part of their high-stakes testing programs. Experts predict that in time, large-scale testing programs will move toward computer-based testing.

As there has been a growing interest in computer-based testing in K–12 large-scale assessments, several comparability studies have involved elementary and secondary students over the past few years. Russell and Haney (1997) investigated the mode effects on middle school students' performances on open-ended items in writing, science, math, and reading, and multiple-choice and short-answer items in language arts, science, and math from the National Assessment of Educational Progress (NAEP), and extended writing test items. They found that the mode effect was not significant for the multiple-choice items, but a substantial effect was found for the open-ended items. The results showed that students who were used to writing on the computer performed better when they responded to the open-ended test using a computer rather than using pencil and paper. Similar findings were found in studies by Russell (1999) and Russell and Plati (2001), who found that students who were accustomed to writing using a computer performed better on the open-ended tests when they wrote using a computer than when they wrote by hand. Pommerich (2004) investigated the item-level mode effects of English, reading, and science reasoning tests in grades 11 and 12 and found that examinees responded differently to some items under the various interface features, although the mode effect in general was small. Pommerich found that the paging condition group outperformed the scrolling condition group in the reading and science reasoning tests, and the automatic scrolling group performed slightly better than the semi-automatic scrolling group in the English test. Two comparability studies on the online versions of the NAEP math and writing tests showed that the paper group significantly outperformed the computer group in the eighth-grade NAEP mathematics test, but no mode effect was found for the eighth-grade NAEP essay test (Sandene, Horkay, Bennett, Allen, Braswell, Kaplan, et al., 2005). The NAEP studies also found that students' familiarity with computers was related to their performance. Particularly, hands-on measures of keyboarding skill were found to be a significant predictor of students' performances on the NAEP online writing test (Sandene, et al., 2005). Although the NAEP studies have directly investigated administration mode effects in the K–12 large-scale assessments, the NAEP is a low-stakes assessment and the lack of consequences for its results could affect student motivation to take the test seriously, and the results may not be generalized to high-stake statewide assessments.

Recently, a number of comparability studies were conducted for statewide computerized assessments. A study of the 2003 Kansas state online assessment in seventh-grade mathematics found no statistically significant differences between paper and computerized versions (Poggio, Glasnapp, Yang, & Poggio, 2005). Nichols and Kirkpatrick (2005) investigated the mode effect for the Florida state assessment in high school reading and mathematics. They found that for both reading and mathematics, mean raw score, mean scale scores, and passing rates were slightly higher for PPT than for CBT, although the mode effect was not significant. Fitzpatrick and Triscari (2005) conducted comparability studies for the Virginia High School End of Course online tests in various subject areas including Earth Science, Biology, and Chemistry. The overall results showed that item parameters and cut score results are comparable across administration modes. Choi and Tinkler (2002) found that the computerized Oregon statewide reading and mathematics tests were more difficult for third graders, but the paper version of the test was more difficult for tenth graders. They also found that mode effects were more pronounced in reading tests than in mathematics tests. Similar findings were reported in the study by Way, Davis, and Fitzpatrick (2006), who investigated the comparability of paper and online versions of the Texas statewide tests in mathematics, reading/English language arts, science and social studies at grades 8 and 11. Overall, the results showed that the tests were more difficult for the online group than for the paper group. The authors also reported that administration mode effects were more evident for ELA than for other subjects. Further item-level analysis on the same tests conducted by Keng, McClarty, and Davis (2006) showed that the mode effect was significant for ELA items with long passages and math items involving graphing and geometric manipulations that required more scrolling through the screen, which supported similar findings by Pommerich (2004). In a recent comparability study of the statewide end-of-course English test, Kim and Huynh (in press-a) found that scores obtained from PPT and CBT were comparable at both the item and test level. A rather large mode difference, however, was found in the reading comprehension section. In another study, Kim and Huynh (in press-b) examined the stability of Rasch latent trait across modes of administration of the statewide grade 9 English test for the total group and subgroups of gender and ethnicity. It was found that the Rasch latent trait remained stable across modes of administration, and the computerized format appeared to have no adverse impact on item-level performance of students as a whole as well as performance of subgroups of students. Although recent studies discussed above have begun to shed some light on administration mode effects, research on effects of administration modes across various academic subject areas in statewide testing programs is still limited. In addition, limited published studies exist on comparability of statewide assessments at the item

level. Therefore, more research is needed to thoroughly investigate various measurement issues concerning comparability of PPT and CBT at both the item and total score level for various academic subject areas in statewide testing programs.

The purpose of this study was to examine comparability of student scores obtained from computerized and paper-and-pencil formats of one Southeastern state's large-scale statewide end-of-course (EOC) examinations in the two subject areas of Algebra and Biology. Research questions of interest were the following:

1. Are there differences in scale scores between CBT and PPT?
2. Do item parameters obtained from CBT and PPT differ from those from the item bank?
3. Are ability estimates at the content domain level similar between CBT and PPT?
4. Are the underlying constructs at the content domain level equivalent between CBT and PPT?

Methods

Participants

Fifteen middle and high schools within five districts in a Southeastern state voluntarily participated in the study. This study used data from students who took the EOC tests in both CBT and PPT modes. Based on the state's 2005 school poverty index (a range of 0–100, with higher indices indicating greater poverty), the range of the participating schools' poverty indices was from 16.5 to 92.4, with an average of 49.5. There were 788 examinees for Algebra and 406 for Biology. Table 1 (next page) reports the demographic characteristics of the examinees for each subject area. It should be noted that the demographic make up of the sample does not necessarily represent the entire middle/high school students in the state. Particularly, African Americans and Hispanics were underrepresented in the sample.

Table 1: Demographic Characteristics of the Students

	Algebra (N = 788)		Biology (N = 406)	
Demographic Characteristic	***n***	**%**	***n***	**%**
Gender				
Female	437	55.5	238	58.6
Male	351	44.5	168	41.4
Ethnicity				
African American	258	32.7	44	10.8
White	479	60.8	332	81.8
Other	51	6.5	30	7.4
Eligible for free or reduced lunch	285	36.2	77	19.0

Instruments

This study used two different parallel test forms with live items for PPT and CBT. In statewide assessment programs, alternate forms are often used for different modes of test administrations because of test security and the need to release student reports on a strict time schedule. Content specialists and psychometricians constructed test forms by selecting items to meet the content specifications and the targeted difficulty level in the operational test blueprints as much as possible. All operational forms were pre-equated to be as equivalent as feasible. The EOC tests are required for all students who enroll in the gateway courses in grades nine through twelve. The tests are weighted only 20 percent in the determination of students' final grades in the course; the scores are not used for passing or failing. Questions on each test are aligned with the state curriculum standards for each course and are designed to assess students' mastery of these standards. The tests are un-timed and standards-based, and composed of all multiple-choice items. There were 50 items for Algebra and 55 items for Biology. The one-parameter Rasch dichotomous model (Rasch, 1960; Wright & Stone, 1979) was used for calibrating multiple-choice items; in placing the field-test items on the item bank scale, the operational item parameters were anchored at the bank difficulty values.

The computerized version of the test developed by a contractor was a fixed-length form, and was delivered via the Internet. The minimum hardware requirements needed to deliver CBT via the Internet were as follows: Pentium II 266 Mhz (for PC), iMAC 233 Mhz (for Apple/Macintosh), 128 MB RAM, 500 MB Available Disk, VGA Display (640 × 480) or Flat-panel Display (800 × 600), and Mouse/Pointing Device. Items were presented individually on a computer screen. For certain items, students needed to scroll down the screen to see an entire question and response options. During the exam, students were provided with a variety of online tools such as a compass, eraser, choice eliminator, calculator, highlighter, periodic table of elements, ruler, and straightedge tools. A practice exam was designed to help students practice using different online tools. Response review was allowed in CBT; that is, students were able to review and change their responses after entering their responses during exam time. At the end of the exam, a review screen was displayed so that students could review which question had been answered, had not been answered, or marked for review before the final submission of their answers.

Procedure

This study used a counter-balanced, repeated measure design in order to control for order effects. Counterbalancing at the individual student level within a classroom was impractical. Therefore, with a list of the classrooms provided by participating schools, staff at the State Department of Education randomly assigned intact classrooms to either (a) a PPT first-CBT second condition or (b) a CBT first-PPT second condition. Participating schools were given one additional testing window week to administer their EOC tests. They were given one week for CBT, one week for PPT, and one week for make-up testing. Three consecutive weeks were selected by participating districts within the state testing window: May 2–June 9, 2005. In order to control for motivation effects, student scores on the first test were not reported to students until they finished the second test. In addition, students were allowed to count a higher score in their final grade. For Algebra, there were 456 students in the PPT first-CBT second and 332 students in the CBT first-PPT second condition. For Biology, there were 156 students in the PPT first-CBT second condition and 250 students in the CBT first-PPT second condition.

Analyses

Differences in Scale Scores

Analyses were conducted to determine if there were differences in student scale scores between CBT and PPT. The distribution of scale score differences was first examined. Then, a two-way repeated measures analysis of variance (ANOVA), with one within-subject factor (mode) and one between-subject factor (order) was performed on scale scores to assess the effects of mode (PPT/CBT) and the mode-by-order interaction. If no mode-by-order interaction effect was found, scores of the two student samples (the PPT given first and CBT given first group) were collapsed for a comparison of scores across administration modes. An effect size measure, g, for dependent groups (Grissom & Kim, 2005, p. 67) was calculated as follows:

$$g = \frac{\overline{Y}_{CBT} - \overline{Y}_{PPT}}{S_p}, \text{ and } S_p = \left[\frac{s^2_{CBT} - s^2_{PPT}}{2^{1/2}} \right]$$

where $\overline{Y}_{CBT}$ is the mean scale score of CBT, $\overline{Y}_{PPT}$ is the mean scale score of PPT, S_p is the pooled standard deviation, s^2_{CBT} is the variance of CBT, and s^2_{PPT} is the variance of the PPT. Statistical analysis was performed using Statistical Package for Social Science 13.0 software (SPSS Inc., 2004).

Differences in Item Parameters

As parallel forms with different sets of items were used for CBT and PPT, a direct comparison in item difficulty parameters between CBT and PPT was not possible. Therefore, item difficulty parameters from each of the administration modes were compared to those from the item bank. The Rasch dichotomous model was used to recalibrate multiple-choice items (Rasch, 1960; Wright & Stone, 1979) using the WINSTEPS program (Linacre, 2005). After rescaling all item parameters to a common scale, the stability of the item parameters was examined by plotting the recalibrated item parameters against the item bank parameters. The robust Z for each item was calculated in order to examine if there was a significant difference between the item bank and new item parameters as follows:

$$\text{Robust Z} = \frac{D - Median}{0.74(IQR)}$$

where D is the difference between the bank (item bank) parameter and new (recalibrated) item parameter, *Median* is the median of the differences, and *IQR* is the interquartile range of the difference. Differences with an absolute robust Z of 1.645 or larger were considered 'significant.'

The robust Z statistic has been used in large-scale assessment programs of states such as Maryland (Maryland State Department of Education, 2005), Minnesota (Minnesota Department of Education, 2004), and South Carolina (South Carolina Department of Education, 2003, p. 48) to detect items that are unstable from one test administration to the next. Next, average absolute difference (AAD) statistics were calculated as follows:

$$\text{AAD} = \frac{\sum_{i=1}^{n} |\hat{b}_i - b_i|}{n}$$

where $\hat{b}_i$ is a recalibrated item parameter for the i^{th} item, b_i is an item bank parameter for the i^{th} item, and n is the total number of items.

In addition to item-level differences, test-level differences were examined by constructing a double axis graph that plots two test characteristic curves (TCCs) of CBT and PPT and the differences of the two TCCs on the same graph. Another double axis graph was constructed for two test information functions (TIFs) of CBT and PPT and the differences of the two TIFs. The expected raw score and test information were computed for each Rasch ability score using SAS 9.1.2.

Differences in Ability Estimates at the Content Domain Level

The analyses were conducted on Rasch ability estimates (i.e., theta) at the content domain level rather than raw scores or scale scores. As aforementioned, a different set of items was used for CBT and PPT, thus using raw scores to compare student performance at the content domain level between CBT and PPT was not appropriate. Given that the operational-test forms were constructed from the item bank, the bank item parameters were used to compute ability estimates. Items on both forms of CBT and PPT were put on a common scale, with all forms covering the same content domains; that is, the CBT and PPT theta estimates for each content domain were on the same theta metric. Given that there was no scale score computed at the content domain level, it was also not practical to use scale scores for the analyses. Even if there were scale scores at this level, the scale scores are just linear transformations of theta, so results would not change if the analysis was done at the scale score level rather than at the theta level. WINSTEPS software (Linacre, 2005) was used for estimating Rasch ability.

The Algebra test had three major content domains and 10 sub-domains: *Understanding Functions* (4 sub-domains), *Linear Functions* (4 sub-domains), and *Quadratic and Other Functions* (2 sub-domains). The Biology test had two major content domains and 12 sub-domains: *Inquiry* (6 sub-domains) and *Biology* (6 sub-domains). Some of the sub-domains contained too few items to conduct the content domain level analysis. After conferring with content experts at the State Department of Education, items from two or more sub-domains within each major domain were merged to define the larger domain as shown in Table 2 (next page). The Rasch ability estimates at the domain level were compared between CBT and PPT for each subject by conducting a repeated measures analysis of variance (ANOVA), with two within-subjects factors (administration mode and content domain) using SPSS 13.0. An administration mode by content domain interaction effect was of primary interest for this research question. Paired *t*-tests were performed when results of the repeated measures ANOVA were significant. Bonferroni adjustment was applied for multiple comparisons.

Table 2: Item Distributions by Content Domain for Algebra and Biology

Subject	Content Domain	Number of Items
Algebra		
	A1. Relationships, Linear Quadratic Functions, Data Representations	10
	A2. Generalizations, Algebra Symbols, Matrices, Algebraic Expressions	10
	A3. Representations, Interpretations	12
	A4. Equations, Inequalities, Linear Equations	10
	A5. Quadratic, Other Functions	8
	Total	**50**
Biology		
	BI. Inquiry	14
	B2. Cell, Matter, Energy, Organization	15 (CBT); 16 (PPT)
	B3. Heredity, Biological Evolution	12
	B4. Interdependence of Organisms, Behavior, Regulation	14 (CBT); 13 (PPT)
	Total	**55**

Equivalence of Construct at the Content Domain Level

The construct equivalence at the content domain level between CBT and PPT was tested by confirmatory factor analyses (CFA) using the LISREL 8.5.1 software (Jöreskog & Sörbom, 2001). The analyses were performed at the content domain level using Rasch ability estimates as the observed variables rather than using raw scores or scale scores for the same reason aforementioned. First, confirmatory factor analyses of the one-factor model were carried out separately for CBT and PPT in order to assess the adequacy of the model for each subject area. In the one-factor model, all domain scores were loaded onto a single factor. Hu and Bentler (1999) suggest that at least two fit indexes should be used simultaneously to reduce discrepancies across fit indexes. Therefore, for the current study, the goodness of fit of the model was tested via the χ^2 statistic, Root Mean

Square Error of Approximation (RMSEA), Standardized Root Mean Square Error Residual (SRMR), Comparative Fit Index (CFI), and Non-Normed Fit Index (NNFI). A non-significant value of χ^2 is indicative of model fit, while a significant value of χ^2 is indicative of model misfit. Based on the recommendations by Hu and Bentler (1999), criterion values for model with a good fit are RMSEA < 0.06, SRMR < 0.08, CFI > 0.95, and NNFI > 0.95.

Second, confirmatory factor analyses of the two-factor model were conducted in order to test measurement invariance between CBT and PPT. In the two-factor model, content domain-level scores from CBT were loaded as Factor 1 and the content domain-level scores from PPT were loaded as Factor 2. Error variances for each pair of the content domain scores between CBT and PPT were correlated in order to reflect the repeated-measures nature of the study. The following steps were used to test three different levels of invariance. First, all parameters for PPT and CBT were freely estimated (Model 1, baseline model). Second, factor loadings were constrained to be equal for each pair of the domains between PPT and CBT (Model 2). Third, factor loadings and error variances were constrained to be equal for each pair of the domains between PPT and CBT (Model 3).

A χ^2 difference test is widely used for testing measurement invariance. A non-significant χ^2 difference supports a higher level of invariance. For instance, a non-significant difference in χ^2 values between Model 1 and Model 2 suggests that factor loadings are invariant between PPT and CBT. Differences in χ^2, however, are sensitive to sample sizes and thus, Cheung and Rensvold (2002) recommend using various goodness-of-fit indexes to test for measurement invariance, rather than the biased χ^2 difference. Specifically, they proposed that when changes in CFI values are smaller than or equal to .01, measurement invariance should not be rejected. Following Cheung and Rensvold's recommendation, the current study used the change in CFI of .01 or less as a criterion for measurement invariance.

Results

Differences in Scale Scores

Figures 1 and 2 display the distributions of score differences between CBT and PPT in Algebra and Biology, respectively. Positive values indicate a higher scale score on CBT than on PPT, while negative values indicate the reverse. For Algebra, there were more negative signs than positive signs, indicating that more students scored higher on PPT than on CBT. Specifically, more than half of the students scored higher on PPT than on CBT. For Biology, there were slightly more positive signs than negative signs. Only four more students scored higher on CBT than on PPT. Slightly over 10% of students received the same scores on PPT and CBT.

Figure 1: Algebra Distribution of Score Differences

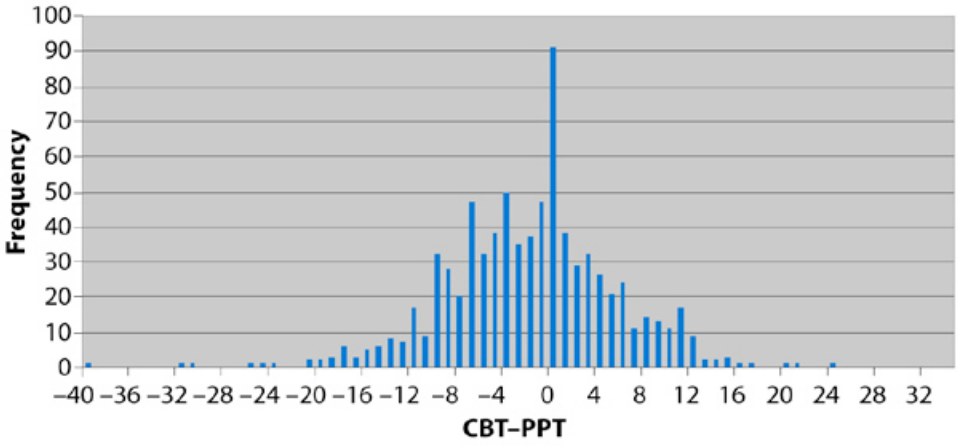
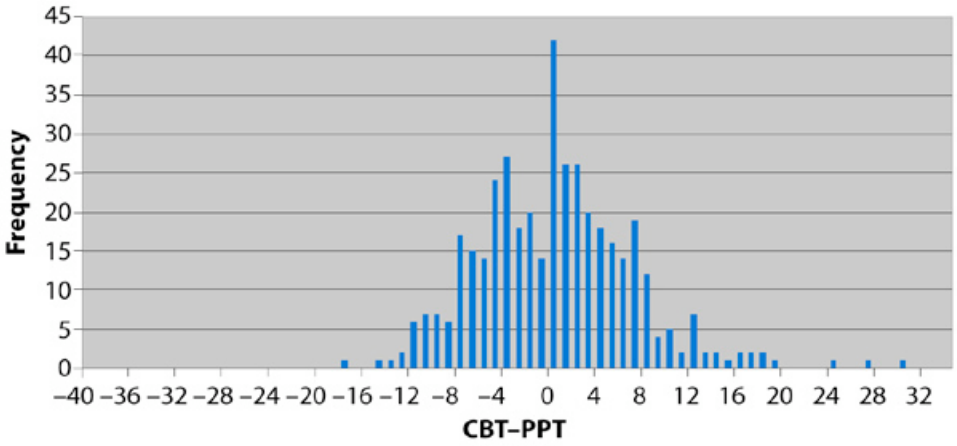


Figure 2: Biology Distribution of Score Differences



For both Algebra and Biology, the two-way repeated measures ANOVA revealed no significant mode-by-order interaction ($F(1, 786) = 4.48$, $p > .01$ for Algebra; $F(1, 404) = 1.33$, $p > .01$ for Biology), suggesting that the two administration order groups can be collapsed for a comparison of mode effect. For Algebra, the mean scale score was higher for PPT ($M = 83.26$, SD = 11.97) than for CBT ($M = 81.27$, $SD = 10.82$) by 1.99 points. The ANOVA result showed that the mode effect was significant ($F(1, 786) = 52.53$, $p < .01$). Using the mean (M) and SD just reported, the (CBT–PPT) significant mean differences were converted into an effect size. The corresponding effect size was 0.17. Using the Cohen's convention (Cohen, 1988, p. 25), this effect size was judged to be either negligible or small. Biology showed a small discrepancy in the mean scale scores between PPT ($M = 78.83$, $SD = 11.76$) and CBT ($M = 78.97$, $SD = 10.98$) and the mode effect was not significant ($F(1, 404) = 0.03$, $p > .01$).

Differences in Item Parameters

Overall, item parameters were quite stable for both subject areas. Figures 3 and 4 (next page) present the scatter plots of the rescaled item and the bank item parameters for PPT and CBT for Algebra and Biology, respectively. A high correlation with the item bank parameters was found for both PPT ($r = .93$ for Algebra; $r = .94$ for Biology) and CBT ($r = .93$ for Algebra; $r = .93$ for Biology). Based on the robust Z statistic, Algebra had three items for PPT and two items for CBT showing a significant difference between the recalibrated item and item bank parameters. For Biology, three items for PPT and five items for CBT showed a significant difference. The AAD for Algebra (.31 for PPT; .37 for CBT) and Biology (.29 for PPT; .33 for CBT) showed that there was a little difference between the recalibrated item and the item bank parameters.

Figure 3: Algebra Scatter Plots of the Rescaled Item and the Bank Item Parameters

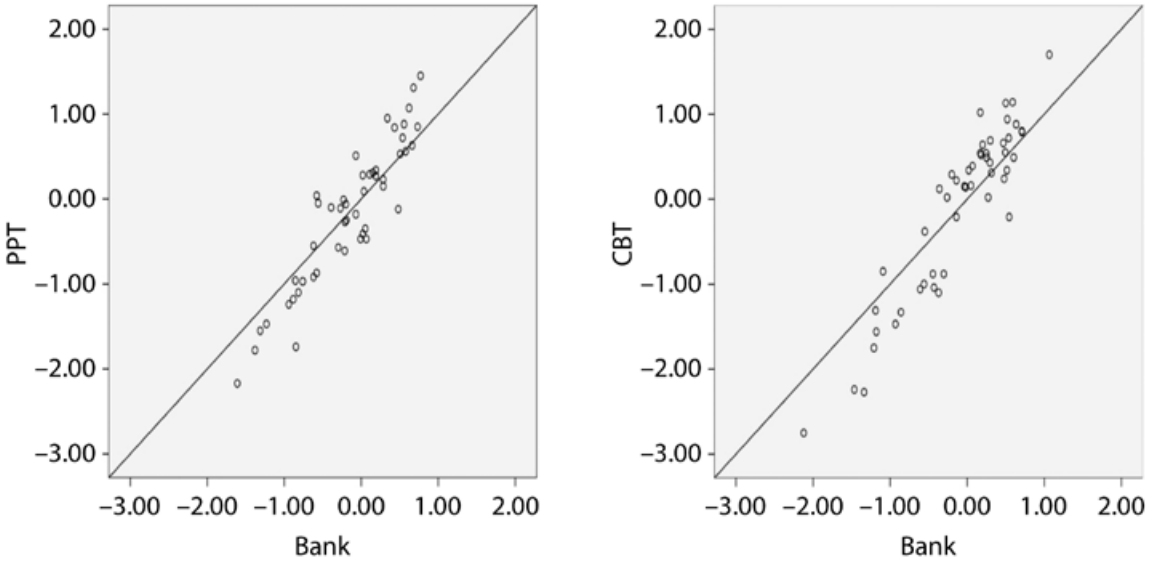
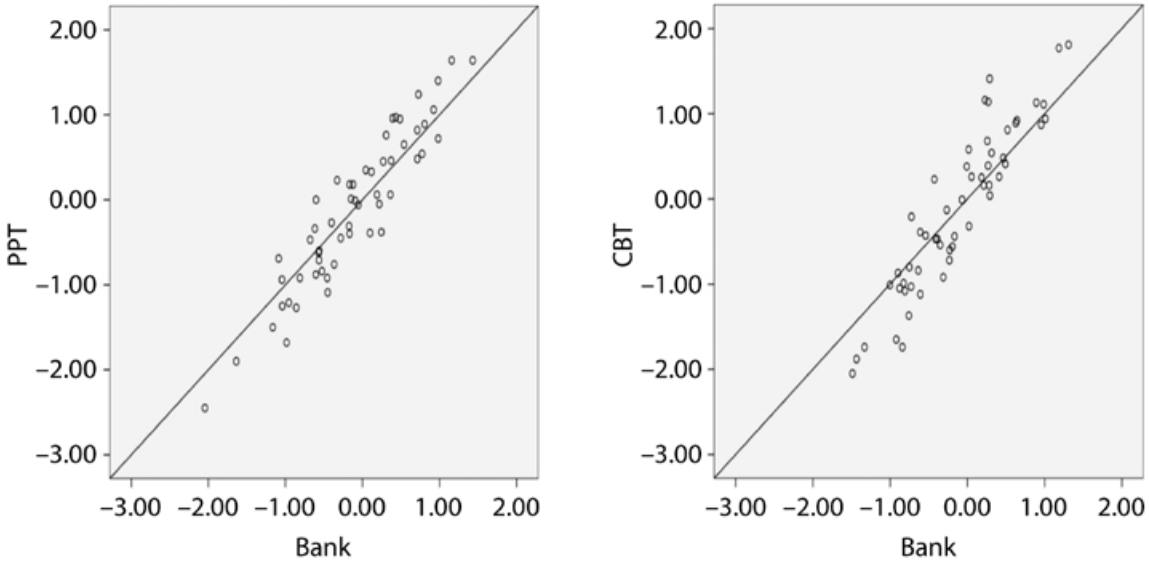


Figure 4: Biology Scatter Plots of the Rescaled Item and the Bank Item Parameters



Figures 5 and 6 (next page) illustrate the two TCCs and the discrepancy of the two TCCs for Algebra and Biology, respectively. The vertical axis at the left side of the plot gives expected raw scores

(i.e., expected number correct scores, $\tau = \sum_{j=1}^{n} P_j(\theta)$).

The vertical axis at the right side of the plot gives the discrepancy of the two TCCs. For both subject areas, the two TCCs appeared to be very close to each other across the entire theta scale. For Algebra (Figure 5, next page), the TCC for PPT was slightly higher than the TCC for CBT at theta = –1.3 and above. That is, at a given level of ability, PPT was easier than CBT. For the rest of the theta scale (theta = –1.4 and below), the reverse was observed, that is, CBT was easier than PPT. Across most parts of the theta scale, differences did not exceed 0.5 unit. A difference of 0.5 unit or larger was observed between theta = –0.5 and theta = 1.4. For Biology (Figure 6, next page), the two TCCs were very close to each other. A difference of less than 0.5 units was observed across the entire theta scale if there was any difference.

Figure 5: Algebra Test Characteristic Curves

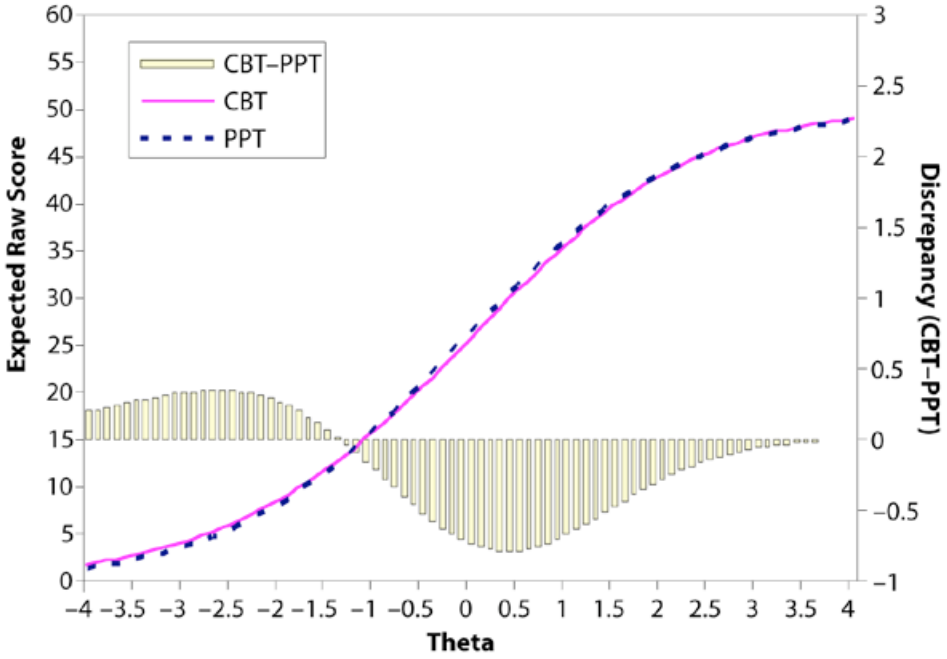
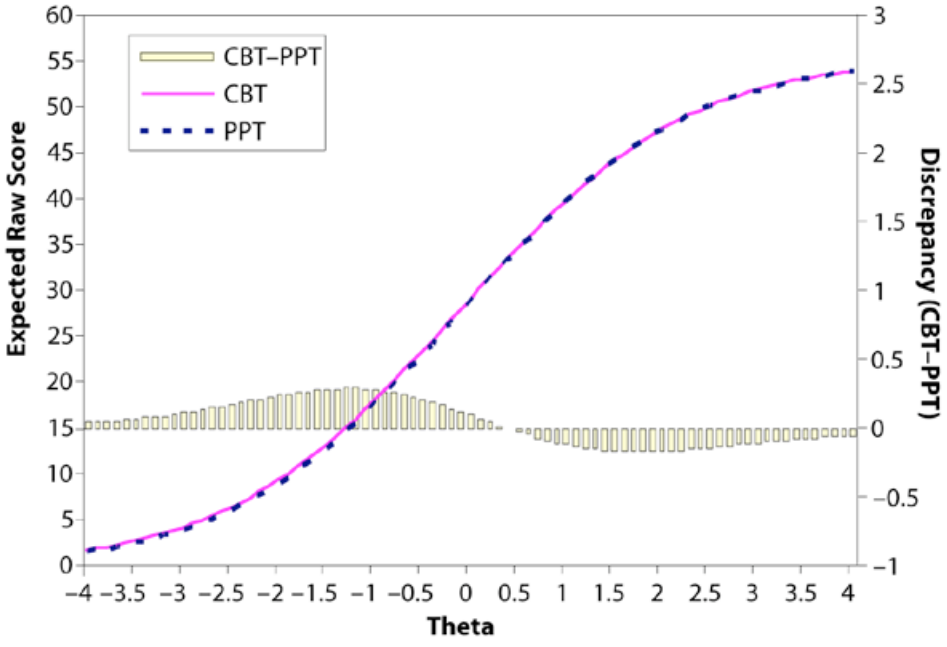


Figure 6: Biology Test Characteristic Curves



Figures 7 and 8 (next page) illustrate the two TIFs and the discrepancy of the two TIFs for Algebra and Biology, respectively. The vertical axis at the left side of the plot gives the test information

$$(\sum_{j=1}^{n} I_j(\theta)).$$

The vertical axis at the right side of the plot gives the discrepancy of the two TIFs. Overall, for both subject areas, two TIFs overlapped closely. For both PPT and CBT, the TIF's maximum point was observed near the center of the theta scale. For Algebra (Figure 7, next page), the TIF for PPT was higher than the TIF for CBT between theta = –.14 and theta = –0.4, with a difference of 0.5 unit or larger, that is, at a given ability level, PPT provided more information than CBT. For the rest of the theta scale, the two TIFs overlapped closely. The ability level corresponding to the maximum test information was similar for PPT (theta = –0.1) and CBT (theta = 0.1). For Biology (Figure 8, next page), the two TIFs overlapped very closely across the entire theta scale. Differences were very small (less than 0.4 unit) if there were any differences. For Biology, maximum test information was observed at theta = –0.1 for both administration modes.

Figure 7: Algebra Test Information Functions

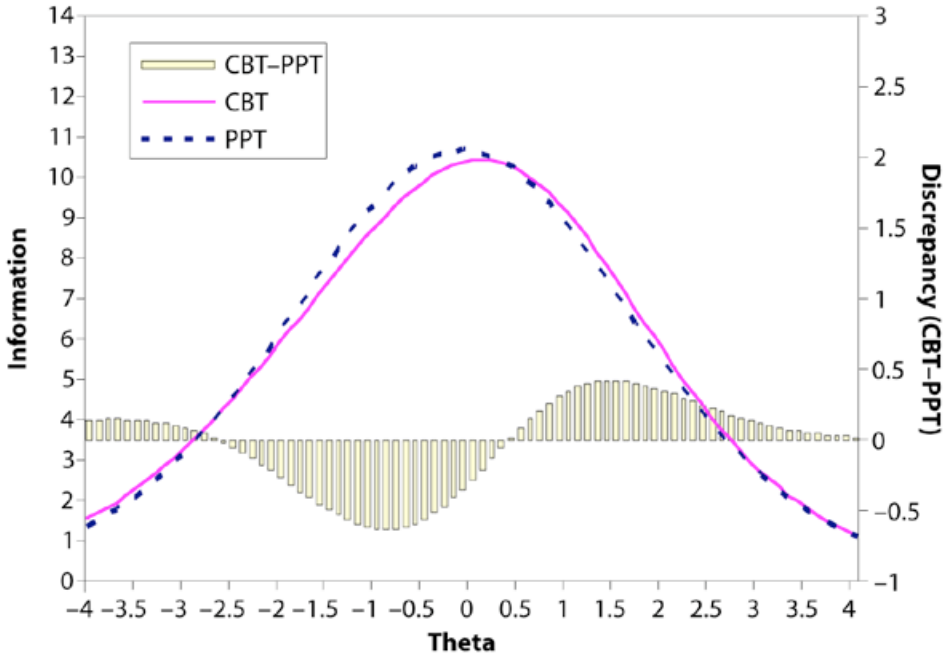
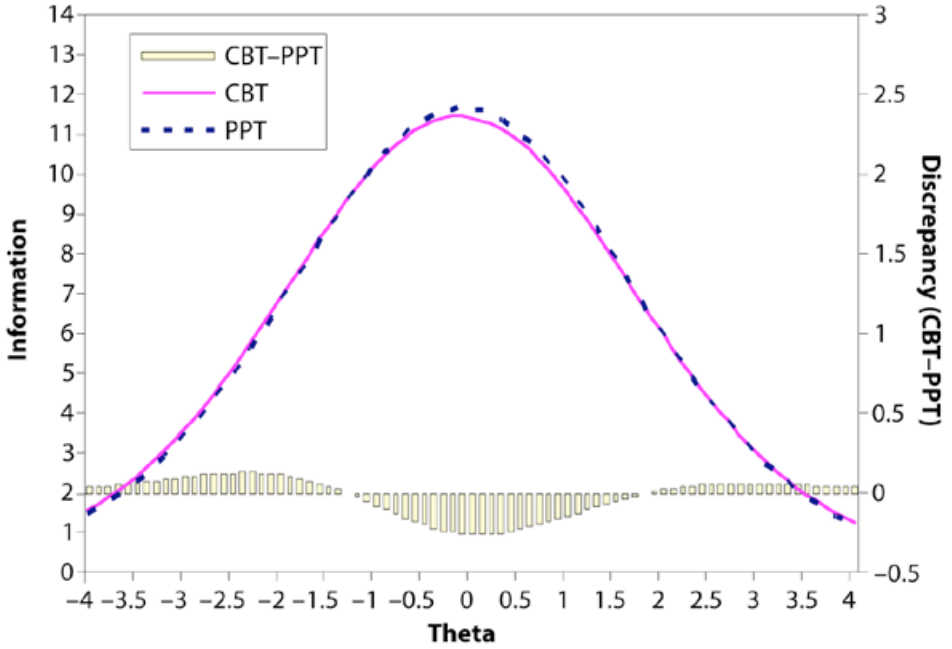


Figure 8: Biology Test Information Functions



Differences in Ability Estimates at the Content Domain Level

Table 3 provides descriptive statistics for the Rasch ability estimates at the content domain level. For Algebra, mean ability estimates of all five domains were higher for PPT than CBT. For both administration modes, the highest mean ability estimate was in domain *A4*, while the lowest mean ability estimate was in domain *A5*. Differences between means of PPT and CBT were small, ranging from 0.12 to 0.25. For Biology, two domains, *B1* and *B4* showed higher mean ability estimates for PPT than for CBT, while the other two domains, *B2* and *B3* showed the reverse finding. Of the four domains, domain B1 showed the highest mean ability estimates for both PPT (0.93) and CBT (0.87).

Table 3: Descriptive Statistics of the Ability Measures for Algebra and Biology

		PPT		CBT	
Subject	**Content Domain**	**Mean**	***SD***	**Mean**	***SD***
Algebra					
	A1. Relationships, Linear Quadratic Functions, Data Representations	0.63	1.18	0.42	1.06
	A2. Generalizations, Algebra Symbols, Matrices, Algebraic Expressions	0.55	1.27	0.35	1.22
	A3. Representations, Interpretations	0.66	1.34	0.41	1.11
	A4. Equations, Inequalities, Linear Equations	0.78	1.46	0.56	1.27
	A5. Quadratic, Other Functions	0.35	1.36	0.23	1.26
Biology					
	Bl. Inquiry	0.93	1.30	0.87	1.20
	B2. Cell, Matter, Energy, Organization	0.55	1.03	0.64	1.05
	B3. Heredity, Biological Evolution	0.58	1.17	0.64	1.12
	B4. Interdependence of Organisms, Behavior, Regulation	0.78	1.27	0.61	1.00

For both subject areas, the sphericity assumption appeared violated based on Mauchley's test of sphericity so that the Huynh-Feldt degrees of freedom adjustment was applied in calculating the significance of all *F* ratios for the repeated factors. The repeated-measures ANOVA for Algebra revealed significant main effects for mode ($F(1, 787) = 77.73$, $p < .01$) and domain ($F(4, 3114) = 41.82$, $p < .01$). The interaction effect of mode × domain was not statistically significant ($F(4, 3091) = 1.52$, $p > .05$). For Biology, the main effect of mode was significant ($F(3, 1206) = 29$, $p < .01$). Importantly, a significant mode by domain interaction effect was observed ($F(3, 1190) = 3.21$, $p < .01$). Given that there was a significant interaction effect of mode by domain, paired *t*-tests were conducted to examine CBT–PPT differences on each of the four content domains in Biology. A statistically significant difference was found for domain *B3 (Heredity & Biological Evolution)* at the .05/4 level (.0125). By convention, (Cohen, 1988, p. 25), the effect size of –0.15 for *B3* was considered small.

Equivalence of Construct at the Content Domain Level

Table 4 presents the results of the one-factor model CFA. Although the χ^2 statistic and the RMSEA indicated that the model did not adequately fit the data for the paper-version of the Biology test, the overall results suggested that the data had an adequate fit to the model for all tests, as indicated by SRMR, NNFI, and CFI.

Table 4: Goodness-of-Fit Indexes of the Model by Subject and Administration Mode

Subject	Mode	*df*	χ^2	SRMR	RMSEA	NNFI	CFI
Algebra	PPT	5	7.61	.01	.03	1.00	1.00
	CBT	5	10.36	.01	.04	.99	1.00
Biology	PPT	2	6.51*	.02	.07	.98	.99
	CBT	2	0.69	.01	.00	1.01	1.00

Note: *$p < .05$; SRMR = standardized root mean square residual, RMSEA = root mean square error of approximation, NNFI = non-normed fit index, CFI = comparative fit index

Results of the series of invariance tests of the two-factor model and fit indexes are summarized in Table 5. For Algebra and Biology, the differences in χ^2 values between Model 1 and Model 2 were statistically significant at $p < .05$ level. However, the ΔCFI of –.01 indicates that equivalence constraints on factor loadings did not adversely affect model fit. The difference in χ^2 values between Model 2 and Model 3 was insignificant for Algebra, but significant for Biology. However, for both subject areas, the ΔCFI was .01, suggesting that equivalence constraints on factor loadings and error variances did not reduce the model fit. In addition, the SRMR, RMSEA, and NNFI met the criteria for a good model fit for all three models.

Table 5: Goodness-of-Fit Indexes for Tests of Invariance for Algebra and Biology

Subject	Mode	*df*	χ^2	$\Delta\chi^2$	CFI	ΔCFI	SRMR	RMSEA	NNFI
Algebra	Model 1	29	45.80	—	1.00	—	.02	.03	.99
	Model 2	33	72.91	27.11	.99	–.01	.04	.04	.99
	Model 3	38	81.93	9.02	.99	0.0	.04	.04	.99
Biology	Model 1	15	21.11	—	1.00	—	.02	.03	.99
	Model 2	18	35.01	13.90	.99	–.01	.05	.05	.99
	Model 3	22	61.79	26.78	.98	–.01	.06	.06	.97

Note: CFI = comparative fit index, SRMR = standardized root mean square residual, RMSEA = root mean square error of approximation, NNFI = non-normed fit index

Summary and Discussion

This study investigated the comparability of student scores obtained from PPT and CBT of large-scale statewide end-of-course examinations in the two subject areas of Algebra and Biology. Student performances were compared at the item-level, subtest (i.e., content domain)-level and whole test-level by examining scale scores, item parameter estimates, test characteristic curves, test information functions, Rasch ability estimates at the content domain level, and the equivalence of the construct being measured. Results of this study found no clear evidence of item-level mode effects in both subject areas. When recalibrated item parameters were compared to the item bank parameters, item parameter estimates were quite stable with a high correlation between the recalibrated and the item bank parameters. Although some individual items appeared to be affected by the mode of administration, overall differences in item parameter estimates and the average absolute difference were fairly small for both subject areas. In addition, both TCCs and TIFs demonstrated similar patterns between PPT and CBT, suggesting the test-level comparability across administration modes.

For Algebra, CBT and PPT were comparable at the content domain level, as evidenced by insignificant interaction effect of mode by domain. For Biology, there was a significant interaction effect of mode by domain. The domain, *Heredity & Biological Evolution*, showed a statistically significant difference between the modes, but the magnitude of the difference was small. Although we do not have a full understanding for this result, this can lead to further investigation by content experts and test developers to identify possible sources of the difference. Results of invariance testing showed that administration modes did not affect the construct being measured. Factor loadings and error variances appeared to be similar between PPT and CBT for both subject areas, suggesting that a similar construct was measured, regardless of the mode of administration.

In sum, results of the current study provide some empirical evidence of comparability of statewide PPT and CBT in Algebra and Biology at the item-level, subtest-level, and whole test-level. The results for Algebra are somewhat consistent with those of Fitzpatrick and Triscari (2005). The findings for Biology also support those of the previous studies on science tests, although fields of science measured in other studies may not be identical to the current study (e.g., Fitzpatrick & Triscari, 2005; Pommerich, 2004). An important limitation of this study involved its reliance on voluntary participation. Schools in which technology was more accessible to students might have been more likely to participate, which could affect the findings. The convenience nature of the sample in this study may cause sampling bias that may limit generalizabilty of findings across

different settings. Future studies should include a more representative sample with diverse backgrounds. Another important research direction includes understanding the role of school characteristics. For example, how will technology-rich school environments impact student performance on computer-based testing? Another limitation of this study is that alternate test forms were used for PPT and CBT. Although alternate test forms are considered equivalent, differences in test forms may be a confounding factor in administration mode effects. Thus, caution should be exercised when interpreting the results of this study. The results of the study are only tentative and reasonable to the extent that the forms were properly equated. Despite the limitations of this study, this study adds to the existing literature because it investigated the administration mode comparability at the item-level and test-level using various analytical methods and thus provided more evidence to support the comparability of computerized and paper formats of the statewide assessments. The findings in this study can provide policymakers and educators with additional information for evaluation of computerized tests in statewide testing programs, and will lead to future policy discussions.

References

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Drasgow, F (2002). The work ahead: A psychometric infrastructure for computerized adaptive tests. In C.N. Mills, M.T. Potenza, J.J. Fremer, & W.C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 67–88). Hillsdale, NJ: Lawrence Erlbaum.

Fitzpatrick, S., & Triscari, R. (2005, April). *Comparability studies of the Virginia computer-delivered tests*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Grissom, R.J. & Kim, J.J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1–55.

Jöreskog, K., & Sörbom, D. (2001). LISREL 8.5.1 for Windows [Computer Software]. Lincolnwood, IL: Scientific Software International, Inc.

Keng, L., McClarty, K.L., & Davis, L.L. (2006, April). *Item-level comparative analysis of online and paper administrations of the Texas Assessment of Knowledge and Skills*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

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

Kim, D.-H., & Huynh, H. (in press-b). Transition from paper-and pencil to computer-based testings: Examining stability of Rasch latent trait across gender and ethnicity. In E.V. Smith, Jr. & G.E. Stone (Eds.), *Applications of Rasch measurement in criterion-reference testing: Practice analysis to score reporting*. Maple Grove, MN: JAM Press.

Linacre, J. M. (2005). WINSTEPS version 3.59.1 [Computer Software]. Chicago, IL: Author.

Maryland State Department of Education (2005). 2005 MSA Reading Technical Report. Retrieved from http://www.marylandpublicschools.org/msde/divisions/planningresultstest/2005+MSA+Reading+Technical+Report, September 2, 2006.

Minnesota Department of Education (2004). Minnesota Basic Skills Tests Technical Manual 2004 Administration. Retrieved from http://education.state.mn.us/mdeprod/groups/Assessment/documents/Report/006696.pdf, May 30, 2007.

Nichols, P., & Kirkpatrick, R. (2005, April). *Comparability of the computer-administered tests with existing paper-and-pencil tests in reading and mathematics tests*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Poggio, J., Glasnapp, D.R., Yang, X., & Poggio, A.J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3*(6). Retrieved July 9, 2005, from http://www.jtla.org.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2*(6). Retrieved July 9, 2005, from http://www.jtla.org.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*. Retrieved September 19, 2004, from http://epaa.asu.edu/epaa/v7n20/.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5*. Retrieved August 5, 2003, from http://epaa.asu.edu/epaa/v5n3.html.

Russell, M., & Plati, T. (2001). Effects of computer versus paper administration of a state-mandated writing assessment. *TC Record.Org*. Retrieved June 19, 2005, from http://www.tcrecord.org/Content.asp?ContentID=10709.

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project, research and development series* (National Center for Education Statistics Publication No. NCES 2005–457). Washington, DC: U.S. Government Printing Office.

SAS Institute Inc. (2004). SAS version 9.1.2 [Computer Software]. Cary, NC: Author.

South Carolina Department of Education (2003). *Technical documentation for the 2003 Palmetto Achievement Challenge Tests of English language arts, mathematics, science, and social studies*. Columbia, SC: Author.

SPSS Inc. (2004). SPSS version 13.0 for Windows [Computer Software]. Chicago, IL: Author.

Swanson, C.B. (2006, May 4). Tracking U.S. Trends. *Education Week's Technology Counts, 25*(35), 50–53.

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

Wright, B.D., & Stone, M.H. (1979). *Best test design*. Chicago, IL: MESA Press.

Author Biographies

Do-Hong Kim is an Assistant Professor at the University of North Carolina at Charlotte, Department of Educational Leadership. Her current research focuses on measurement issues concerning statewide computer-based assessments and large-scale testing of students with disabilities. She can be contacted at dkim15@uncc.edu

Huynh Huynh is a Professor of Educational Studies and of Statistics at the University of South Carolina. He is a Fellow of the American Statistical Association. He served as an Associate Editor of Psychometrika and of Journal of Educational Statistics, a joint publication of the American Statistical Association and American Educational Research Association. He can be contacted at hhuynh@gwm.sc.edu

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College