Research Report

Evaluating the Comparability of Paper-and-Pencil and Computerized Versions of a Large-Scale Certification Test

Gautam Puhan

Keith A. Boughton

Sooyeon Kim

Research & Development

October 2005
RR-05-21

Evaluating the Comparability of Paper-and-Pencil and Computerized Versions of a Large-Scale Certification Test

Gautam Puhan, Keith A. Boughton, and Sooyeon Kim
ETS, Princeton, NJ

October 2005

Abstract

The study evaluated the comparability of two versions of a teacher certification test: a paper-and-pencil test (PPT) and computer-based test (CBT). Standardized mean difference (SMD) and differential item functioning (DIF) analyses were used as measures of comparability at the test and item levels, respectively. Results indicated that effect sizes derived from the SMD were small ($d < 0.20$) and not statistically significant ($p > 0.05$), suggesting no substantial difference between the two test versions. Moreover, DIF analysis revealed that reading and mathematics items were comparable for both versions. However, five writing items were flagged for DIF. Substantive reviews failed to identify format differences that could explain the performance differences, so the causes of DIF could not be identified.

Key words: PPT, CBT, differential item functioning, item impact, standardized mean difference

Acknowledgements

The authors would like to thank Tim Davey, Cathy Wendler, Anna Kubiak, and Sandip Sinharay for their helpful comments and suggestions.

Perspectives/Theoretical Framework

The effectiveness of achievement tests as tools that yield scores that can be validly interpreted regardless of the mode of delivery of these tests (e.g., paper and pencil vs. computer) is often questioned (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999). For example, scores derived from computer-based tests (CBT) as compared to paper-and-pencil tests (PPT) might reflect not only the examinee’s proficiency in the construct being measured but also the level of computer proficiency. This affects the construct being measured and disrupts the comparison and interpretation of test scores across the two modes of administration.

Many testing programs are increasingly administering the same test in both PPT and CBT formats. For example, the TOEFL® program concurrently delivers PPTs and CBTs in approximately 228 countries every year. Similarly, the GRE® General test is offered in the United States, Canada, and many other countries in paper-and pencil- and computer-based formats. More recently, the PRAXIS™ program started administering the Pre-Professional Skills Tests (PPST®) via CBT in addition to PPT. Mills, Potenza, Fremer, and Ward (2002) speculated that this trend will continue to grow because of an increase in availability of microcomputers in educational settings, a substantial improvement in the speed of computers, and a significant reduction in cost.

CBTs have many advantages over PPTs, which may include faster score reporting, savings on paper and personnel resources and costs of scoring services (Wise & Plake, 1990), and development of new methods of assessment such as simple adaptations of multiple-choice items to more innovative item types (Jodoin, 2003). Despite these advantages, an important question that arises when tests are administered in both formats is whether or not the scores produced are interchangeable (Wang & Kolen, 2001; Gallagher, Bridgeman, & Cahalan, 2002). For example, scores derived from CBTs as compared to PPTs might reflect not only the examinee’s proficiency on the construct being measured but also differences in formatting (including typing versus handwriting) and/or computer proficiency.

There is a large body of research that documents the comparability of scores obtained from PPTs and CBTs. Mazzeo and Harvey (1988) in their review of studies comparing PPTs and CBTs indicated that CBTs tended to be more difficult than PPT versions of the same tests. Similarly, a meta-analysis conducted by Mead and Drasgow (1993) suggested that the constructs being measured across the two modes were similar for power tests but not for speeded tests. Finally, a more recent study by Gallagher et al. (2002) found that performance across PPT and CBT versions of tests differed for subgroups based on gender and ethnicity. For example, female test takers performed poorly on CBTs whereas African-American and Hispanic test takers benefited from this format. However, other researchers have found that PPTs and their CBT counterparts yield comparable scores. For example, Taylor, Jamieson, Eignor, and Kirsch (1998) studied the comparability of PPTs and CBTs for the 1996 administration of the TOEFL and found no meaningful difference in performance for examinees taking the two different versions. Similarly, Wise, Barnes, Harvey, and Plake (1989) contended that PPT and CBT versions of achievement tests yield very similar scores.

Since the results of these studies are inconsistent and the use of computers have become commonplace, it is even more important to examine whether scores obtained from the two different modes of delivery are in fact comparable (Gallagher et al., 2002). In addition, the bulk of these studies have focused on mean differences in test performance across PPTs versus CBTs. However, in order to gain more precise understanding about the nature of differences between modes, item-level performance (e.g., differential item functioning or DIF) should also be assessed. For example, items based on longer reading passages displayed on multiple screens in CBT as compared to the same passages presented on a single page in PPTs may lead to differential performance on these items (Thompson, Thurlow, & Moore, 2002). Also, in the case of constructed response items, it has been found that raters are more lenient toward handwritten essays as compared to computer typed essays. One reason for this occurrence, found by Arnold et al. (1990), was that raters gave students the benefit of the doubt in situations where the handwriting became difficult to read.

Since fairness is an important concern in the field of educational measurement, it is important to ensure that scores obtained from both PPTs and CBTs are comparable and thus measure the same construct. Hence, the purpose of this study is to compare the performance of examinees who took the PPT version of a teacher certification test with another group that took the same form in CBT. It should be noted that there was one overall difference between items in PPT and CBT formats for the full test. The items in PPT format had the multiple-choice options clearly marked as A, B, C, D, or no error on the passages, and the examinees were instructed to choose the correct option. However, in the CBT format, these options were not marked as A, B, C, D, or no error. Instead the options were presented as underlined texts in the passages, and the examinees were instructed to click on the underlined text that denoted the correct option. This difference was assumed to have no effect on the examinees' ability to respond to an item in the PPT or CBT formats. In the United States, these tests hold extremely high stakes because test-takers who do not pass these tests are not eligible to enroll in teaching programs and then to teach in those states that require a passing score on these tests. Therefore, it becomes especially important to ensure that these tests do not unfairly favor one group of test-takers over another based on whether they took the test in PPT or CBT format.

Method

Teacher Certification Test and Sample

This study used test data collected in a 2003 administration of a large-scale teacher certification test administered in 31 states. This test measures basic proficiency in reading, writing, and mathematics and is used for entrance into teaching programs. The paper-and-pencil and computerized versions of the reading and mathematics tests have 40 items each and the writing test has 45 items. In addition to the number of items stated above, the computerized versions of the reading, writing, and mathematics tests had approximately five additional items that are used for pre-testing and are not used in the final scoring. The paper-based tests in reading and mathematics are each 60-minute multiple-choice tests. The writing test includes a 30-minute multiple-choice section and a 30-minute essay section. The computer-based tests in reading and mathematics are each 75-minute multiple-choice tests. The writing test includes a 45-minute multiple-choice section and a 30-minute essay section. The increase in time for the reading and mathematics sections and the multiple-choice section of the writing test in the computer-based tests is due to an increase in the total number of items resulting from the additional pre-test items and also to allow for tutorials and the collection of background information from test-takers.

Six groups of examinees classified by mode of administration (i.e., PPT vs. CBT) and content area (i.e., reading, writing, and mathematics) were analyzed. It should be noted that the examinees were free to choose between either a PPT or CBT version of the test and therefore, there was no random assignment of examinees to either version of these tests. This is important to note because performance differences found in PPT and CBT versions of tests may be due to actual ability differences in the test-taking populations (i.e., test impact) rather than differences in testing format (PPT vs. CBT).

The sample sizes differed slightly for the analyses conducted at test and item levels. At the test level, there were 1,122 examinees in reading, 1,050 examinees in writing, and 1,136 examinees in mathematics for the CBTs. An equal number of examinees were used for the PPTs in reading, writing, and mathematics, respectively. These examinees were sampled from a larger population in the PPT group based on propensity score matching (a more detailed discussion on propensity score matching will follow in the analytical procedure section). For the item-level analysis, there were 1,122 examinees in reading, 1,050 examinees in writing, and 1,136 examinees in mathematics for the CBTs. For the PPTs, the study used two random subsamples of 2,000 examinees from the larger population in reading, writing, and mathematics, respectively. Thus, the item-level analysis involved two separate runs. For the first run, all available examinees for the CBTs and a random subsample of 2,000 examinees for the PPTs were used. In the second run, the exact same analysis was replicated with the second random subsample of 2,000 examinees for the PPTs.

This study was broken into three steps:

1. The standardized mean difference across the mode (i.e., PPT vs. CBT) of the test was evaluated in order to identify the overall difference in performance at the test level.
2. Mode DIF analyses were conducted for the PPT and CBT versions of the tests to identify items that may function differentially across the two modes at the item level.
3. Substantive analysis of the items flagged as DIF was conducted by test reviewers in order to identify the sources of mode DIF.

Analytical Procedure

At the *test level*, comparability of PPTs and CBTs was evaluated using the standardized mean difference or SMD (see Gallagher et al., 2002). SMD reports mean differences in terms of standard deviation units thereby establishing a common metric for comparing performance for examinees who took the paper-and-pencil or computerized versions of the tests. The SMD is calculated as

$$d = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{SD_{Pooled}} \text{ and } SD_{Pooled} = \sqrt{\frac{s_1^2\left(n_1 - 1\right) + s_2^2\left(n_2 - 1\right)}{n_1 + n_2 - 2}},$$

where d is the effect size, $\overline{X}_{G1}$ is the mean of the PPT group, $\overline{X}_{G2}$ is the mean of the CBT group, and SD_{Pooled} is the pooled standard deviation of the PPT and CBT groups. According to Cohen (1988), the following guidelines for defining the d statistic are meaningful: (a) small effect when d is approximately 0.20; (b) moderate effect when d is approximately 0.50; and (c) large effect when d is approximately 0.80. Also, all effect sizes were tested for statistical significance by converting the d statistic into a t-value (where $d = 2t/\sqrt{(df)}$) and checking whether this value was greater than the critical t-value in the t-distribution.

As mentioned earlier, the examinees were free to choose between either a PPT or CBT version of the test, and there was no random assignment of examinees to either version of these tests. Therefore, a simple comparison of performance between the PPT and CBT groups at the test level can be misleading in that such a comparison may not reveal the effect of mode of delivery (PPT vs. CBT) on test performance per se because results can be confounded by other factors that lead these examinees to choose a particular testing format. To overcome this potential problem, the analysis used *propensity scores* (Rosenbaum & Rubin, 1983) to match PPT and CBT examinees on a single variable (i.e., the propensity score). This tended to balance any consistent differences in the distributions of these groups. The propensity score was estimated using a logistic regression model where the dependent variable was testing mode (PPT or CBT) and the independent variables were gender, language, test repeater status, race, GPA, and educational level. The empirically estimated regression weights from the logistic regression were used to compute the propensity score for each examinee in the PPT and CBT groups, separately. The propensity score was a weighted sum of the variables in the logistic regression and is denoted as:

$$Y = b_1x_1 + b_2x_2 + b_3x_3 + ... + b_nx_n,$$

where $x_1, x_2, ...x_n$ represent the examinee’s values of the selected variables and $b_1, b_2, ...b_n$ represent the weights determined by the logistic regression.

After the propensity scores were calculated for each examinee in the CBT and PPT groups, the PPT group examinees were sampled on propensity scores to match propensity scores for the examinees in the CBT group, thereby leading to two distributions (i.e., PPT vs. CBT) that were demographically similar.

At the *item level*, comparability of PPTs and CBTs was evaluated using the DIF detection procedure SIBTEST (Shealy & Stout, 1993) to identify items that function differentially for the two versions of the tests. A variety of statistical procedures for detecting DIF have been developed (Camilli & Shepard, 1994; Millsap & Everson, 1993; Potenza & Dorans, 1995). Of these, the Mantel-Haenszel (MH), simultaneous item bias test (SIBTEST), and logistic regression (LR) have been the most commonly used (e.g., Allalouf, Hambleton, & Sireci, 1999; Gierl & Khaliq, 2001; Gierl, Rogers, & Klinger, 1999). Moreover, of these procedures, SIBTEST has been found to be more effective than MH and LR in detecting DIF (Jiang & Stout, 1998; Bolt & Stout, 1996; Gierl et al., 1999). SIBTEST has two well-documented benefits. First, SIBTEST uses a regression estimate of the true score instead of the observed score to match students with the same ability. As a result, examinees are matched on a latent rather than an observed score. Second, SIBTEST can be used to assess DIF iteratively by initially using all the items from the matching test and systematically removing DIF items from the matching test until a subtest of items without DIF is identified (Shealy & Stout, 1993). Furthermore, Gierl et al. (1999) have shown that SIBTEST identifies more DIF items as compared to MH and LR. From a test-development point of view, detection of more DIF items may be problematic because of the enormous costs incurred to develop test items. However, from an interpretation point of view, identification of more DIF items may result in a more thorough analysis of the test items leading to a more comprehensive test interpretation. Therefore, SIBTEST was used in the present study to identify items with DIF.

DIF analysis was conducted using SIBTEST in which each item was used as a studied item and the remaining items were used as the matching subtest. SIBTEST provides an overall statistical test and a measure of the effect size ($\hat{B}_{UNI}$) for each item. In the SIBTEST framework, DIF is conceptualized as a difference between the probabilities of selecting a correct response for examinees with the same levels of the latent attribute of interest (θ). This difference, when found, is attributable to different amounts of nuisance abilities (η) that influence the item response patterns.

The statistical hypothesis tested by SIBTEST is:

$$H_0\text{: } B(T) = P_R(T) - P_F(T) = 0$$

versus

$$H_1\text{: } B(T) = P_R(T) - P_F(T) \neq 0,$$

where *B (T)* is the difference in probability of a correct response on the studied item for examinees in the reference (or advantaged) and focal (or disadvantaged) groups matched on true score; P_R (T) is the probability of a correct response on the studied item for examinees in the reference group with true score T; and $P_F(T)$ is the probability of a correct response on the studied item for examinees in the focal group with true score *T.* According to Roussos and Stout (1996, p. 220) the following $\hat{B}_{UNI}$ values are used for classifying DIF: (a) negligible or Level A DIF: Null hypothesis is rejected and $|\hat{B}_{UNI}| < 0.059$; (b) moderate or Level B DIF: Null hypothesis is rejected and $0.059 \leq |\hat{B}_{UNI}| < 0.088$; and (c) large or Level C DIF: Null hypothesis is rejected and $|\hat{B}_{UNI}| \geq 0.088$. These guidelines were used to classify DIF items in the present study. Also, in all analyses, an alpha level of 0.05 was used with a nondirectional hypothesis test.

An initial concern before conducting the DIF analysis was whether the total scores from the two testing modes (PPT vs. CBT) were sufficiently comparable so that they conveyed the same meaning in the two testing modes. This was important to ensure because the two groups used in the DIF comparison are not equivalent groups and if the total scores did not convey the same meaning across the two testing modes, then an equating adjustment would be necessary. The SMD analysis conducted on matched groups at the test level showed that there was no difference in mean performance in the two testing modes (see Table 1). Furthermore, the PPT and CBT score distributions for the matched groups were tested using the *Kolmogorov-Smirnov* (KS) test to determine whether the two distributions were significantly different for the reading, writing, and mathematics tests. The maximum difference (also know as *D*) for the PPT and CBT score distributions for the reading, writing, and mathematics tests are 0.03, 0.05, and 0.04, respectively. These values are not statistically significant ($p > 0.01$), suggesting that there is no statistically significant difference in the PPT and CBT score distributions. Thus, it seemed that a particular raw score conveyed the same meaning across the two testing modes, and therefore, the DIF analysis could be conducted using unmatched groups.

For the substantive analyses, two test reviewers independently examined the test items flagged as DIF by SIBTEST. Items with a C-level rating on both random subsample runs or a C-level rating for one random subsample run and a B-level rating for the other random subsample run were examined. This decision seems justified since C-level items are typically scrutinized for bias in test reviews (Zieky, 1993). The two test reviewers worked independently and generated substantive explanations about the causes of possible mode DIF. Once the independent reviews were completed, the two test reviewers met to discuss their decisions and reach consensus on the items where they disagreed.

Results

Statistical and substantive analyses were conducted at the test and item levels to evaluate the comparability of the paper-and-pencil and computerized versions of the teacher certification tests. For the *statistical analysis*, comparability of the paper-and-pencil and computerized versions of the tests was evaluated using the standardized mean difference for the two versions of the tests. Additionally, items on the paper-and-pencil and computerized versions of the tests were evaluated for DIF using SIBTEST across the two versions of the tests. For the *substantive analysis*, two test reviewers examined the items that were statistically flagged as DIF and generated substantive interpretations regarding the cause of DIF on those items.

Statistical Analysis

The SMD (i.e., *d* statistic) was calculated for scores obtained from the paper-and-pencil and computerized administrations of the reading, writing, and mathematics tests. The means and standard deviations for the paper-and-pencil and computerized administrations for the three tests are similar (see Table 1). As seen in Table 1, the effect sizes for the PPT and CBT comparisons for the reading, writing, and mathematics tests were small (< than 0.20). Following Cohen's (1988) criteria, these small effect sizes indicate that the PPT and CBT versions of the reading, writing, and mathematics tests are comparable. Furthermore, the *d* statistic computed for the paper-and-pencil versus computerized versions of the tests were not statistically significant ($p >$ 0.05) for the reading, writing, and mathematics tests, suggesting that the paper-and-pencil and computerized versions of these tests do not show a statistically significant difference.

Table 1

Means and Standard Deviations and Effect Sizes for the Computerized and Paper-and-Pencil Versions of the Teacher Certification Test

	Reading		Writing		Mathematics	
	CBT	PPT	CBT	PPT	CBT	PPT
Mean	30.597	30.404	30.877	30.714	27.862	28.402
SD	6.719	7.207	6.911	7.221	7.502	7.729
SMD	0.028		0.023		-0.071	

Note. SMD = standardized mean difference. Statistical significance of the SMD was calculated using an unpaired *t*-test.
**p* < 0.05.

Furthermore, DIF analysis of the reading, writing, and mathematics tests were conducted to identify items that functioned differentially between the PPT and CBT versions of these tests. An initial concern was the possibility of contamination of the matching subtest if a large number of DIF items was found in these tests. To overcome the problem, an iterative approach was used in which a single item DIF analysis was first conducted and items displaying large DIF were removed from the matching subtest and the data was re-analyzed using the purer matching subtest (i.e., DIF free), thereby leading to a more stable set of results. It should also be noted that an initial DIF analysis for the writing test was conducted using all 45 items, in which the last five items showed DIF against the PPT group. Since these five items were also identified as speeded for the PPT version of the test in an earlier study (see Boughton, Yamamoto, & Larkin, 2003) they were dropped from the current analysis, and the final DIF analysis for the writing test was conducted with 40 instead of 45 items.

For the reading test there were no items identified as showing DIF for both PPT and CBT versions of the test. For the writing test, five items (items 1, 11, 15, 24, & 32) were identified as showing DIF. Of these five items, four items favored examinees who took the PPT, and one item favored examinees who took the CBT (see Figure 1 and Table 2). For the mathematics test, there were no items identified as showing DIF. Results of the statistical analyses suggest that performance across the PPT and CBT versions of the reading, writing, and mathematics tests are comparable at the test level (i.e., SMD results). However, DIF analyses suggest that item level differences exist across the PPT and CBT versions of the writing test. Mode DIF was more prominent in the writing test as compared to the reading and mathematics tests in which there were no C-Level DIF items.

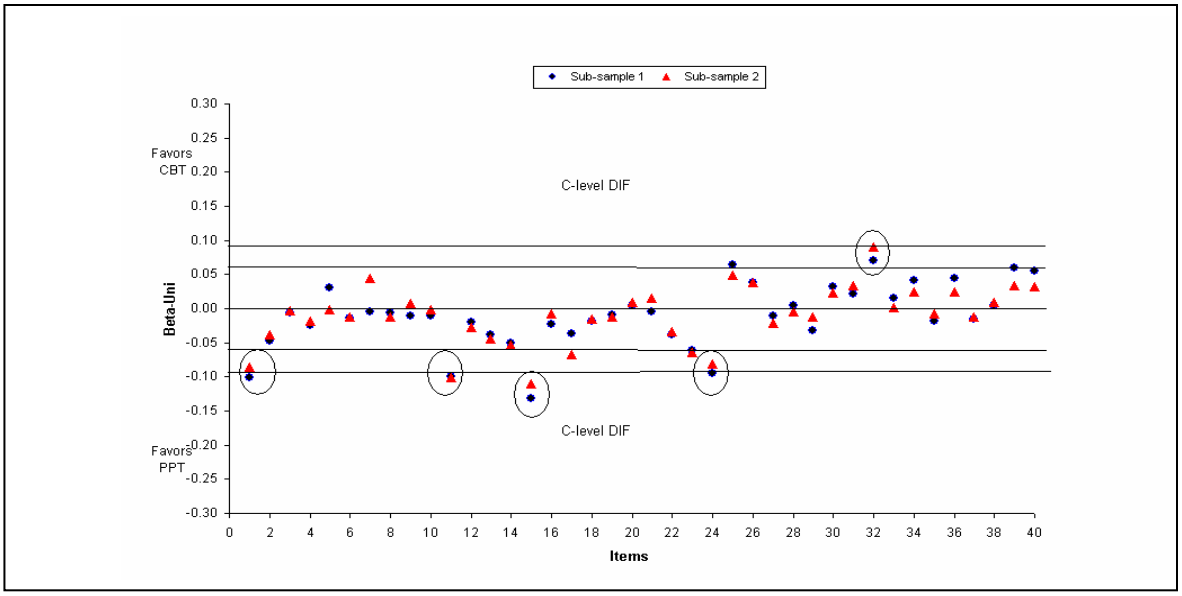

Figure 1. **DIF items for paper-and-pencil and computerized versions of the writing test.**

Table 2

Results of SIBTEST DIF Analysis for the Computerized and Paper-and-Pencil Versions of the Teacher Certification Test: Writing

Item	Writing			
	Random subsample I		Random subsample II	
	$\hat{\beta}_{UNI}$	Level	$\hat{\beta}_{UNI}$	Level
1	-0.101*	C	-0.086*	B
11	-0.099*	C	-0.101*	C
15	-0.131*	C	-0.110*	C
24	-0.095*	C	-0.081*	B
32	0.070*	B	0.090*	C

Note. A negative $\hat{\beta}_{UNI}$ favors paper and pencil test-takers.

$*p < 0.05$.

Substantive Analysis

Statistical methods are useful in detecting DIF items; however, to understand the nature of mode DIF, two test reviewers provided substantive reviews for the items identified as DIF (see, for example, Camilli & Shepard, 1994, p. xiii; Gierl & Khaliq, 2001). Without substantive analysis, it would be difficult to know whether an item that is flagged as DIF is due to differences in the mode of administration of these tests (e.g., paper and pencil vs. computer) or due to actual ability differences between the examinees from the two populations (i.e., item impact). The test reviewers did not find any difference in the items flagged as DIF that may cause differential performance on these items across PPT and CBT formats. Also, since the SMD showed no difference in mean performance across the two testing modes and the KS-test showed no difference in the PPT and CBT score distributions, it is less likely that item impact can be the cause of DIF on these items. Therefore the cause of DIF on these items could not be identified. The test reviewers found an overall difference between items in PPT and CBT formats for the full test but they did not ascribe this difference to be a cause of DIF for items that were flagged statistically. The items in PPT format had the options clearly marked as A, B, C, D, or no error on the passages. However, in the CBT format, these options were presented as blank spaces in the passages. Because this difference was present for the remainder of the items in the test it would be erroneous to conclude that this difference resulted in DIF for the five items that were flagged statistically but not for the remainder of the items on the test.

Discussion and Conclusion

The purpose of the present study was to evaluate the comparability of PPT and CBT versions of a teacher certification test designed to measure basic proficiency in reading, writing, and mathematics. The standardized mean difference (SMD) was used to evaluate the comparability of the PPT and CBT versions of the reading, writing, and mathematics tests. The *d* statistic calculated using SMD was not statistically significant for the reading, writing, and mathematics tests. The effect size measure was also used to obtain a more practical estimation of the magnitude of difference between the PPT and CBT versions of the tests. Evaluation of the effect sizes suggested that these tests were comparable across the PPT and CBT formats. It should be noted, however, that statistically significant results are not always practically important and vice versa. In the current study, although the difference between the standardized means for the PPT and CBT groups were small statistically, it may still affect the pass/fail status of a large number of examinees especially when the total samples are large. Therefore, considerable thought by the testing program must go into determining the practical implications of these results.

Furthermore, the items were analyzed using SIBTEST to identify items that function differentially for examinees who took the tests in PPT or CBT formats. There were no items flagged as DIF for the reading and mathematics tests. There were five items (four favoring PPT and one favoring CBT) that were identified as DIF for the writing test across the two random subsamples that were analyzed. However, substantive reviews failed to identify any difference in format that could lead to DIF on these items. The authors suggest that the testing program should monitor these five items in future testing administrations, and if they continue to show DIF statistically, then they should be replaced with new items.

This study was important, as it has been demonstrated by earlier research that administering tests in PPT and CBT formats may affect the comparability of scores obtained from these testing formats. Since the teacher certification tests have extremely high stakes outcomes for test-takers, it was important to examine whether the tests yielded comparable scores when administered in PPT or CBT formats. Furthermore, by examining item-level performance in addition to test level performance, this study provided an opportunity to review format differences at the item level. As Gallagher et al. (2002) pointed out, with an increase in familiarity of students with computers, an overall measure of difference in test performance due to change in mode of delivery may appear less meaningful today. Thus, it was important to use both statistical and substantive analyses at the test and item level in order to ensure that tests are fair and valid for all, regardless of mode of presentation. As evident, the findings of this study were positive and suggested that the CBT and PPT versions of this teacher certification test are comparable.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Allalouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of translation DIF on verbal items. *Journal of Educational Measurement*, *36*, 185–198.

Arnold, V., Legas, J., Obler, S., Pacheco, M. A., Russell, C., & Umbdenstock, L. (1990). *Do students get higher scores on their word processed papers? A study of bias in scoring hand-written vs. word-processed essays.* Unpublished manuscript.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23,* 67–95.

Boughton, K. A., Yamamoto, K., & Larkin, K. (2003). Modeling differential speededness using a HYBRID psychometric approach. Unpublished manuscript.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Newbury Park: Sage.

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

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. *Journal of Educational Measurement, 39*(2), 133–147.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement*, *38,* 164–187.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). *Consistency between statistical procedures and content reviews for identifying translation DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.

Jiang, H., & Stout, W. (1998). Improved type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioural Statistics*, *23* (4), 291–322.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement, 40*(1), 1–15.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from conventional and automated educational and psychological tests: A review of literature* (College Board Rep. No. 88–8).New York: College Board.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3)*,* 449–58.

Mills, C. N., Potenza, M.T., Fremer, J.J., & Ward, C. W. (2002). *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297–334.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation*. Applied Psychological Measurement*, *19*, 23–37.

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

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, *33*, 215–230.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, *58*, 159–194.

Taylor, C., Jamieson, J., Eignor, D. R., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks* (ETS RR-98-08). Princeton, NJ: ETS.

Thompson, S., Thurlow, M., & Moore, M. (2002). *Using computer-based tests with students with disabilities* (Policy Directions No. 15). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38*(1), 19–49.

Wise, S. L., Barnes, L. B., Harvey, A. L., & Plake, B. S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. *Applied Measurement in Education, 2,* 235–241.

Wise, S. L., & Plake, B. S. (1990). Computer-based testing in higher education. *Measurement and Evaluation in Counseling and Development, 23*(1), 3–10.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.