A Study to Evaluate the Comparability of Paper and Online Versions of the Grade 8 Science Field-Test as part of Arizona's Instrument to Measure Standards (AIMS)

Report prepared for
The Arizona Department of Education

September 30, 2007

PEARSON

A Study to Evaluate the Comparability of Paper and Online Versions of the AIMS Grade 8 Science Field-Test

Abstract

The purpose of this study was to evaluate the comparability between online and paper field-test results for the AIMS grade 8 science field-test administration. The results suggested very similar performance between the paper and online groups participating in the field test. The mean ability estimates for the overall samples of paper and online students matched on previous score and demographic variables were nearly identical. Some minor differences in mean abilities for online and matched paper groups were found across field-test forms and subgroups, but none of these differences were of practical significance. Comparisons of p-values using standardized difference statistics flagged a small set of items exhibiting “significant” mode differences, but the differences went in both directions, that is, some favored students testing by paper and others favored students testing online. In summary, the field-test performance of the paper and online samples was sufficiently comparable to support the future administration of the operational grade 8 science test in both paper and online modes. In addition, the results of the study support equating and reporting scores for the operational test without regard to testing mode.

A Study to Evaluate the Comparability of Paper and Online Versions of the AIMS Grade 8 Science Field-Test

Introduction

Arizona's Instrument to Measure Standards (AIMS) is a Standards-Based test that provides educators and the public with valuable information regarding the progress of Arizona's students toward mastering Arizona's reading, writing and mathematics Standards. In 2007, the Arizona Department of Education (ADE) introduced a grade 8 science field-test as part of the AIMS administration. This field test included administration both by the traditional paper-and-pencil format and by a computer-administered, online format.

When tests are offered in both paper and online formats, professional testing standards indicate the need to evaluate the comparability of test results across the two modes of administration (APA, 1986; AERA, APA, NCME, 1999, Standard 4.10). For this reason, and because ADE is interested in administering the operational grade 8 science test online in the future, Pearson has conducted a study evaluating the comparability of paper and online student performance based on the grade 8 science field-test results.

The comparability of test scores based on online versus paper testing has been studied for more than 20 years. Reviews of the comparability literature research were reported by Mazzeo and Harvey (1988), who reported mixed results, and Mead and Drasgow (1993), who concluded that there were essentially no differences in examinee scores by mode-of-administration for power tests. Paek (2005) provided a summary of more recent comparability research and concluded that, in general, computer and paper versions of traditional multiple-choice tests are comparable across grades and academic subjects. However, when tests are timed, differential speededness can lead to mode effects. For example, a recent study by Ito and Sykes (2004) reported significantly lower performance on timed web-based norm-referenced tests at grades 4-12 compared with paper versions. These differences seemed to occur because students needed more time on the web-based test than they did on the paper test. Pommerich (2004) reported evidence of mode differences due to differential speededness in tests given at grades 11 and 12, but in her study online performance on questions near the end of several tests was *higher* than paper performance on these same items. She hypothesized that students who are rushed for time might actually benefit from testing online because the computer makes it easier to respond and move quickly from item to item.

A number of studies have suggested that no mode differences can be expected when individual test items can be presented within a single screen (Poggio, Glassnapp, Yang, & Poggio, 2005; Hetter, Segall & Bloxom, 1997; Bergstrom, 1992; Spray, Ackerman, Reckase, & Carlson, 1989). However, when items are associated with text that requires scrolling, such as is typically the case with reading tests, studies have indicated lower performance for students testing online (Way, Davis & Fitzpatrick, 2006; O'Malley, 2005; Pommerich, 2004; Bridgeman, Lennon, & Jackenthal, 2003; Choi & Tinkler, 2002; Bergstrom, 1992).

The purpose of this study was to evaluate the comparability between online and paper field-test results for the AIMS grade 8 science field-test administration. During this administration, the same field-test forms were administered both in paper and online formats. This permitted comparisons to be made both at the overall student performance level and at the individual item level. Results from both types of analyses will be presented in this report.

Methodology

We utilized a matched groups design for the comparability study (Way, Um, Lin, & McClarty, 2007; Way, Davis, & Fitzpatrick, 2006). The matched groups approach is really a quasi-experimental method in which comparisons between online and paper groups that have not been sampled to be equivalent are made possible by matching the groups on external variables, such as previous test scores and demographic characteristics. In this design, the same test form or forms are typically administered to the online and paper groups (although this is not required). The advantage of this design is that there is minimum burden on districts and schools because there is no need to assign students to conditions. That is, the online group is compared with a matched sub-sample of the students that take the paper version of the test. The weakness of the design is that the quality of the matching depends upon the relationship of the external variable with the test scores being compared. Pearson has had success with this design using matching variables such as test scores from other subjects administered concurrently or scores from the previous spring's test. In the case of the grade 8 science field-test administration, we used scores on the AIMS reading, mathematics, and writing tests that were also administered in spring 2007.

Although these tests measure different skills, our experience with previous studies suggested that the relationship between performance on these test scores and performance on the science field-tests would be strong enough to support their use in the matched groups analyses.

Data and Psychometric Model

The data for the comparability study were collected from the spring 2007 AIMS administration and included scored (0 or 1) item responses for each participating grade 8 student on one of five randomly-spiraled field-test forms, and scale scores on the operational reading, mathematics, and writing AIMS tests. In addition, we used gender, ethnicity[1], and the field-test form administered as matching variables for the study.

The comparability analyses were carried using item response theory (IRT). Specifically, we used the three-parameter logistic (3PL) model and the MULTILOG program version 7.0.3 (Thissen, Chen & Bock. 2003). We first calibrated the paper-and-pencil science field-test data using MULTILOG, obtaining item and ability parameter (theta) estimates. Next, we did a second run of MULTILOG with the online field-test data, fixing the item parameters at the values obtained from the paper-and-pencil calibrations and estimating student abilities only. To provide more accurate ability estimates for students at extremely low and high proficiencies, we utilized maximum a posteriori (MAP) ability estimation in MULTILOG.

Table 1 summarizes numbers of students taking each of the five field-test forms in paper and online formats and presents descriptive statistics for the estimated thetas and raw scores. There were 11,395 students in the paper-and-pencil group and 6,181 students in the online group included in the study. The sample sizes by form ranged from 2,187 to 2,342 for the paper group and from 1,199 to 1,254 for the online group.

As the data in Table 1 indicate, the paper group had higher mean estimated thetas and higher mean raw scores than the online group overall and for field-tests 1 to 4. For field-test 5, however, the mean theta and raw score was higher for the online group than for the paper group. In general, the performance differences between the two groups were small.

[1] To guard against inadequately-sized matching groups, ethnicity was collapsed into White, Hispanic, and “Other”.

Table 1. Summary Statistics for Science Field-Test Theta Estimates and Raw Scores

		Theta				Raw Score			
Form	**N**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Online									
1	1253	0.04	0.89	-2.14	2.32	19.42	6.41	2	38
2	1254	0.01	0.85	-2.04	2.64	19.13	6.16	4	37
3	1241	-0.04	0.88	-2.28	2.63	23.32	7.27	1	40
4	1234	0.00	0.82	-2.15	2.25	19.85	6.54	2	38
5	1199	0.09	0.86	-2.02	2.53	21.73	7.26	3	40
Overall	6181	0.02	0.86	-2.28	2.64	20.68	6.92	1	40
Paper									
1	2342	0.05	0.91	-2.26	2.99	19.63	6.36	3	41
2	2317	0.05	0.89	-2.14	2.70	19.63	6.43	3	38
3	2333	0.03	0.92	-2.36	2.63	23.76	7.36	4	40
4	2216	0.05	0.88	-2.14	2.94	20.10	6.85	2	38
5	2187	0.05	0.87	-2.10	2.28	21.22	7.20	5	39
Overall	11395	0.05	0.90	-2.36	2.99	20.87	7.03	2	41

Table 2 presents univariate statistics for the scale scores on the AIMS reading, mathematics, and writing tests obtained by students participating in the field-test comparability study. These data also suggest slightly higher performance for the students who took the science field-test forms on paper as compared to those students who took the science field-test forms in the online format. It should be noted that all of the data summarized in Table 2 are based on the operational paper-and-pencil administration of these AIMS tests.

Table 2. AIMS Scale Scores for the Online and Paper Science Field-Test Samples

		Reading				Math				Writing			
Form	**N**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Group Taking Science Field-Tests Online													
1	1253	520.07	56.75	374	800	554.80	58.61	428	800	549.83	66.47	300	800
2	1254	517.38	53.98	390	686	554.43	59.26	410	800	548.77	68.54	300	800
3	1241	519.55	56.38	382	800	556.04	60.48	422	800	545.81	67.55	300	800
4	1234	521.23	56.15	382	800	556.51	61.01	422	800	548.14	68.00	300	800
5	1199	522.12	55.93	390	800	558.12	61.45	403	800	549.49	66.37	300	800
Overall	6181	520.05	55.85	374	800	555.96	60.15	403	800	548.41	67.39	300	800
Group Taking Science Field-Tests on Paper													
1	2342	519.82	55.53	382	800	557.12	60.52	417	800	549.49	68.45	300	800
2	2317	520.25	57.83	365	800	557.01	61.62	410	800	552.12	66.76	300	800
3	2333	521.70	57.52	382	800	557.64	60.47	417	800	551.72	67.36	300	800
4	2216	521.19	57.74	374	800	557.54	61.07	428	800	553.12	67.63	300	800
5	2187	522.32	56.77	382	800	557.74	59.39	422	800	552.28	66.57	300	800
Overall	11395	521.04	57.08	365	800	557.40	60.62	410	800	551.72	67.36	300	800

Table 3 summarizes the intercorrelations among the science field-test theta estimates and the AIMS reading, mathematics, and writing test scale scores. As would be expected, the three operational tests are moderately-to-highly correlated with the science field-test theta estimates. In addition, the patterns of the intercorrelations are very similar for the paper and online groups.

Table 3. Intercorrelations among Science Field-Test Theta Estimates and Reading, Math, and Writing Scale Scores for the Online and Paper Samples

Online					Paper				
		Reading	**Math**	**Writing**			**Reading**	**Math**	**Writing**
Overall	**Theta**	0.77	0.73	0.58	**Overall**	**Theta**	0.78	0.75	0.57
	Reading		0.79	0.64		**Reading**		0.78	0.64
	Math			0.59		**Math**			0.59
Form 1	**Theta**	0.77	0.71	0.57	**Form 1**	**Theta**	0.78	0.74	0.57
	Reading		0.76	0.65		**Reading**		0.78	0.64
	Math			0.56		**Math**			0.58
Form 2	**Theta**	0.76	0.72	0.57	**Form 2**	**Theta**	0.77	0.74	0.57
	Reading		0.80	0.65		**Reading**		0.8	0.65
	Math			0.60		**Math**			0.60
Form 3	**Theta**	0.78	0.74	0.59	**Form 3**	**Theta**	0.78	0.77	0.59
	Reading		0.79	0.64		**Reading**		0.78	0.64
	Math			0.59		**Math**			0.59
Form 4	**Theta**	0.77	0.73	0.58	**Form 4**	**Theta**	0.79	0.76	0.57
	Reading		0.8	0.65		**Reading**		0.78	0.64
	Math			0.61		**Math**			0.59
Form 5	**Theta**	0.80	0.73	0.58	**Form 5**	**Theta**	0.78	0.73	0.56
	Reading		0.79	0.62		**Reading**		0.78	0.62
	Math			0.59		**Math**			0.58

Matched Sampling Procedures

To implement the matched samples approach, we considered the operational AIMS scale scores, gender, ethnicity, and the assigned field-test form as matching variables for the comparability analyses. To simplify the matching process, we first used multiple regression procedures to produce a composite score variable according to the following procedures:

1) Using the students who took the paper-and-pencil field tests, we regressed their reading, mathematics, and writing scale scores on their science field-test ability estimates.

$$\hat{\theta}_{predicted\_theta} = \beta_0 + \beta_1 X_{1(\mathrm{Re}ading\_SS)} + \beta_2 X_{2(Math\_SS)} + \beta_3 X_{3(Writing\_SS)} .$$

2) The resulting regression weights were applied to all students (paper and online) to obtain a predicted theta for each student.
3) All students (paper and online groups together) were then broken into 15 groups based on their predicted theta values.
4) This resulted in a 15 (previous score groups) by 5 (field-test forms) by 3 (ethnic groups) by 2 (gender groups) grid that was used in the matched sampling.
5) To improve optimal matching, students with missing values on any of the matching variables were dropped from the study.

The next step in the analyses was to randomly sample students from the online field-test group and match them to samples of students from the paper field-test group with identical profiles of composite scale score group, gender, ethnic group, and field-test form. We used a bootstrap sampling approach and replicated the sampling and matching process for 100 iterations. Within each bootstrap iteration, we sampled (with replacement) 3,000 students from the total online sample. For each of these online students, we randomly sampled a "matching" student from the paper field-test sample with the same score group, gender, ethnic group, and field-test form, also with replacement. Sampling with replacement across replications permitted us to estimate bootstrap standard errors to assist in interpreting differences between the online and paper ability estimates (see Kolen & Brennan, 2004, p. 232-235, for a discussion of bootstrap standard errors).

Comparability Analyses

The bootstrap approach permitted us to compare the paper and online groups within each of the 100 iterations and to aggregate the results over iterations. We compared the groups in terms of ability estimates and item-level performance. We also examined differences by gender and ethnic group within each field-test form and overall across all forms. Finally, we examined score differences in terms of an IRT-based scale equating that might be made to "correct" for any mode effects based on the match samples approach. Each of these analyses is described in more detail below.

The mean differences in theta estimates (and the mean effect size, see Cohen [1992]) between the online and paper testing modes were calculated across all students and for each subgroup. A standardized difference was calculated for each ‘matched’ comparison using the following equation:

$$Zdif = \frac{\overline{D}_{Diff}}{\sqrt{SE^{2}_{Diff}}}$$

where $\overline{D}_{Diff}$ is the grand mean of the differences between mean online and mean paper theta estimates over the 100 replications and SE_{diff} and is the bootstrap standard deviation of the differences over the replications.

The effect size between two group means at each replication was calculated by the following equation:

$$EffectSize = \frac{\overline{X}_{Group1} - \overline{X}_{Group2}}{\sqrt{\frac{(SD^{2}_{Group1} + SD^{2}_{Group2})}{2}}}$$

Overall effect sizes for the theta estimates (and individual item p-values) were then calculated based on the averages of the effect sizes over replications.

To examine possible IRT-based scaling adjustments that might be made to correct for mode effects, we utilized mean/sigma (M/S) transformations. Assuming that the matched sampling approach resulted in equivalent online and paper testing groups, student theta estimates from the two modes would be expected to differ only if a mode effect existed. To the extent that the estimates in the two groups were not equivalent, an M/S transformation could be used to adjust for the mode differences using the following equation:

$$\theta_{paper,i} = A\theta_{online,i} + B$$

where $\theta_{online,i}$ and $\theta_{paper,i}$ are the original online and transformed (to the paper scale) theta estimates and A and B are slope and intercept constants derived using the following equations:

$$A = \frac{\sigma(\theta_{paper})}{\sigma(\theta_{online})}$$

$$B = \hat{\mu}(\theta_{paper}) - A\hat{\mu}(\theta_{online})$$

where $\sigma(\theta_{online})$ and $\sigma(\theta_{paper})$ are the standard deviations of the theta estimates for the online and matched paper samples, and $\hat{\mu}(\theta_{paper})$ and $\hat{\mu}(\theta_{online})$ are the mean of the theta estimates for the online and matched paper samples. Conceptually, if no testing mode effect exists, the A constant should equal "1" and the B constant should equal "0". To the extent that the estimated A and B constants are close to 1 and 0, the paper and online results can be considered comparable.

Within each bootstrap replication, we calculated A and B transformation constants, both for the overall group and for each field-test form. We then obtained overall A and B transformation constants by averaging these constants over the 100 replications.

Results

The primary results of the study are summarized in Table 4, which presents the mean theta estimates for the online and matched paper groups and the mean estimated theta differences between the two groups across the 100 bootstrap replications. These results are presented for all students and separately by gender and ethnic groups across the five field test forms combined, and separately for each field-test form. The results presented include mean differences, standard deviations of the differences, standardized differences, and effect sizes as described above. Also presented are the average sample sizes of the comparison groups across the 100 replications. In each replication, there were 3,000 students sampled from each group and the number of students in each category of each matching variable was the same for each group. However, across replications, the sample sizes for each matching category differed slightly; hence, Table 4 lists the "average N" over replications.

Table 4. Summary of Estimated Theta Differences Based on the Matched Sampling Results

Group	N	Mean Theta		Mean Diff.	SD of Diff.	Standardized Difference	Effect Size
		Online	Paper				
All Field-Test Forms Combined							
All Students	3000.00	0.02	0.02	0.00	0.01	-0.20	0.00
White	1192.12	0.46	0.43	0.03	0.02	1.35	0.04
Hispanic	1323.38	-0.29	-0.27	-0.02	0.02	-0.94	-0.02
Others	484.50	-0.21	-0.17	-0.04	0.03	-1.45	-0.05
Female	1493.49	0.03	0.04	0.00	0.02	-0.20	0.00
Male	1506.51	0.01	0.01	0.00	0.02	-0.08	0.00
Field-Test Form 1							
All Students	612.13	0.03	0.02	0.01	0.03	0.36	0.01
White	243.44	0.48	0.45	0.04	0.05	0.71	0.04
Hispanic	265.06	-0.26	-0.26	-0.01	0.04	-0.15	-0.01
Others	103.63	-0.27	-0.27	0.00	0.07	-0.02	0.00
Female	323.12	0.02	0.02	0.00	0.04	-0.04	0.00
Male	289.01	0.05	0.02	0.03	0.05	0.54	0.03
Field-Test Form 2							
All Students	607.14	0.01	0.03	-0.02	0.03	-0.56	-0.02
White	250.02	0.47	0.41	0.06	0.05	1.23	0.07
Hispanic	255.61	-0.34	-0.26	-0.08	0.04	-1.88	-0.11
Others	101.51	-0.20	-0.16	-0.04	0.07	-0.59	-0.05
Female	313.19	0.04	0.05	-0.01	0.04	-0.27	-0.01
Male	293.95	-0.01	0.01	-0.02	0.04	-0.53	-0.03
Field-Test Form 3							
All Students	599.32	-0.04	-0.01	-0.03	0.03	-1.00	-0.03
White	227.52	0.41	0.41	0.00	0.05	-0.07	0.00
Hispanic	276.02	-0.36	-0.34	-0.02	0.04	-0.57	-0.03
Others	95.78	-0.21	-0.09	-0.12	0.06	-1.90	-0.14
Female	294.40	-0.01	0.04	-0.05	0.04	-1.18	-0.06
Male	304.92	-0.07	-0.06	-0.01	0.05	-0.27	-0.01
Field-Test Form 4							
All Students	602.24	0.00	0.03	-0.03	0.02	-1.26	-0.04
White	239.03	0.39	0.42	-0.03	0.04	-0.70	-0.04
Hispanic	269.27	-0.24	-0.23	-0.01	0.04	-0.23	-0.01
Others	93.94	-0.27	-0.17	-0.09	0.07	-1.32	-0.12
Female	291.49	0.03	0.03	0.00	0.04	-0.01	0.00
Male	310.75	-0.02	0.04	-0.06	0.03	-1.69	-0.07
Field-Test Form 5							
All Students	579.17	0.10	0.04	0.06	0.03	2.00*	0.06
White	232.11	0.54	0.45	0.08	0.04	2.05*	0.11
Hispanic	257.42	-0.24	-0.27	0.03	0.05	0.65	0.04
Others	89.64	-0.10	-0.14	0.05	0.07	0.64	0.06
Female	271.29	0.10	0.06	0.05	0.04	1.10	0.06
Male	307.88	0.09	0.03	0.06	0.04	1.65	0.07

* The standardized difference is greater than or equal to 1.96.

Overall, the results in Table 4 suggest comparable science field-test results between the paper and online samples. Across all forms combined and all students, the mean difference in theta estimates was 0.00, the standardized difference was -0.20, and the effect size was 0.00. The comparisons within field-test forms indicated similar evidence of comparability, with the possible exception of Form 5, for which the standardized mean difference for all students and for white students each exceeded 1.96. The largest effect sizes were -0.11 for the Form 2 Hispanic group comparison, -0.14 for the Form 3 “Others” group comparison, and 0.11 for the Form 5 White group comparison-- all within the traditional classification of “small” assigned for effect sizes less than 0.20.

One minor trend in the results seen in Table 4 is that the mean differences for White students tended to be slightly positive (suggesting slightly higher average performance online as compared with paper), whereas, the mean differences for Hispanic and “Other” students tended to be slightly negative (suggesting slightly lower average performance on paper as compared with online). However, where this trend occurred, none of the mean differences for these groups differed significantly from zero.

The scaling constants to transform the online theta estimates to the scale of the paper estimates, averaged over the 100 matched sampling iterations, are shown below. Constants were calculated based on all of the estimated thetas and for the estimated thetas within each of the five field-test forms.

Form	***A***	***B***
Overall	1.0324	0.0018
1	1.0061	-0.0115
2	1.0241	0.0164
3	1.0414	0.0315
4	1.0733	0.0300
5	1.0283	-0.0583

In all cases (overall and within each form), the A constant was greater than 1.0. The B constant was virtually zero overall, and ranged from slightly negative to slightly positive across field-test forms. Based on the A constant, it would appear that the online estimated thetas were less variable than the paper estimated thetas. Figure 1 illustrates this by graphing the differences between the estimated online thetas and the transformed estimated online thetas as a function of the estimated online thetas.

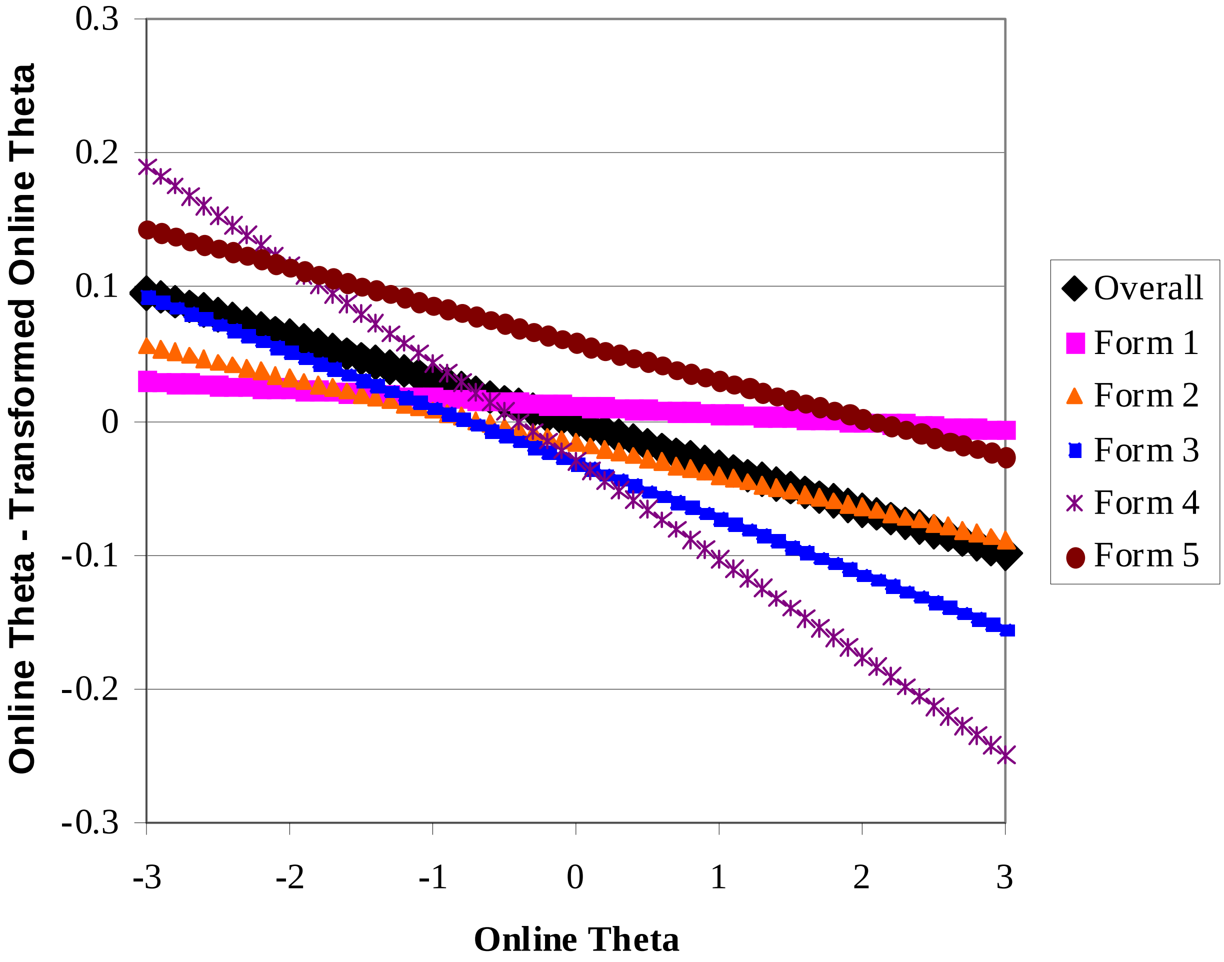

Figure 1. Differences Between Online Theta Values and Transformed Online Theta Values as a Function of Online Theta Values

As Figure 1 indicates, an online theta of -3 is higher than the corresponding transformed online theta, and an online theta of +3 is lower than the corresponding transformed online theta. The patterns of the relationships differ by form, as would be expected from the A and B constants shown above. The graph of the differences based on the overall transformation constants confirm the previous evidence of comparability in that the differences between the original and transformed online thetas are less than 0.10 over the range of estimated thetas.

The results of the p-value comparisons based on the matched samples are presented in Appendix 2 through Appendix 6 by field-test form, and include the online and paper p-values, the differences between p-values, standardized difference statistics, and effect sizes. Items for which the standardized differences exceeded ±1.96 are denoted with an asterisk in the column labeled “Sig.”. There were 24 items flagged for extreme standardized differences across the five forms. Fifteen of the flagged items were easier on paper and nine were easier online. The absolute difference in p-values for the flagged items ranged from 0.04 to 0.10. For all of these flagged items, the effect sizes were 0.20 or less.

It would be useful for content experts to look at the items flagged with extreme p-value differences between the paper and online modes and to consider possible explanations for these differences. Previous research has suggested that items requiring scrolling and items involving graphics may be associated with lower performance when administered online than when administered by paper. Such characteristics may or may not be present in the items flagged in this study. In addition, some consideration could be given to whether individual items flagged as exhibiting mode effects based on this study should be used in the construction of operation test forms. Such consideration might be similar to that given to items flagged as exhibiting differential item functioning (DIF) between subgroups, in that DIF results are typically considered along with other issues, such as the availability of alternate items measuring the same content standards and whether a viable construct-related hypothesis for the DIF can be found.

Discussion and Conclusions

The purpose of this study was to evaluate the comparability between online and paper field-test results for the AIMS grade 8 science field-test administration. The results suggested very similar performance between the paper and online groups. The mean ability estimates for the overall samples of paper and online students matched on previous score and demographic variables were nearly identical. Some minor differences in mean abilities for the online and matched paper groups were found across field-test forms and subgroups, but none these differences were of practical significance. Comparisons of p-values using standardized difference statistics flagged a small set of items exhibiting “significant” mode differences, but the differences went in both directions, that is, some favored students testing by paper and others favored students testing online. In summary, the field-test performance of the paper and online samples was sufficiently comparable to support the future administration of the operational grade 8 science test in both paper and online modes. In addition, the results of the study support equating and reporting scores for the operational test without regard to testing mode.

Although the evidence obtained in this study strongly supports comparability, it should be noted that the results were based on a quasi-experimental approach, and therefore, the potential limitations inherent in this type of design apply. In addition, the data for the study were obtained from a field-test administration in which students may not have been as motivated to perform well as they would be in an operational administration. This could limit the generalization of these results to the operational setting. Finally, despite the positive comparability findings, it is recommended that Arizona encourage schools to use practice tests and online tutorials to prepare students for operational testing by computer, and that schools give some consideration to allowing students to opt out of online testing if they are not comfortable with it. These strategies will help to ensure the comparability between paper and online results as the operational AIMS grade 8 science test is implemented.

References

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessments (APA) (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.

American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1999). S*tandards for educational and psychological testing*. Washington, DC: AERA.

Bergstrom, B. (1992, April). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis*. Paper presented at the annual meeting of the American Educational Research Association: San Francisco.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (ETS RR-01-23). Princeton, NJ: Educational Testing Service.

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. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Hetter, R. D., Segall, D. O. & Bloxom, B. M. (1994). A comparison of item calibration media in computerized adaptive testing. *Applied Psychological Measurement, 18*(3), 197-204.

Ito, K., & Sykes, R. C. (2004). *Comparability of Scores from Norm-Referenced Paper-and-Pencil and Web-Based Linear Tests for Grades 4 – 12*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices* (2nd ed.). New York: Springer.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests. A review of the literature* (ETS RR-88-21). Princeton, NJ: Educational Testing Service.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449-458.

O'Malley, K. J., Kirkpatrick, R., Sherwood, W., Burdick, H. J., Hsieh, M.C., & Sanford, E.E. (2005, April). *Comparability of a Paper Based and Computer Based Reading Test in Early Elementary Grades*. Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada.

Paek, P. (2005). *Recent trends in comparability studies* (PEM Research Report 05-05). Available from http://www.pearsonedmeasurement.com/downloads/research/RR_05_05.pdf.

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). Available 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). Available from http://www.jtla.org.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement*, *26*, 261-271.

Thissen, D., Chen, W-H., & Bock, R. D. (2003). *MUTILOG for Windows, Version 7* [Computer Software]. Lincolnwood, IL: Scientific Software International.

Way, W. D., Um, K., Lin, C., & McClarty, K. L. (2007, April). *An Evaluation of a Matched Samples Method for Assessing the Comparability of Online and Paper Test Performance*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

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.

APPENDICES

Appendix 1. Item Parameter Estimates from the Paper Test Form MULTILOG Calibrations

Note: The following items are excluded from the list.

Excluded Items		
Form	**Item**	**Reason**
3	16	negative *a*
3	39	negative *a*
4	11	negative *a*
5	23	too large *a*
5	36	negative *a*

Form	Item	*a*	*b*	*c*
1	1	0.73	-0.97	0.18
1	2	1.01	-1.25	0.07
1	3	0.69	1.02	0.18
1	4	0.74	0.66	0.15
1	5	1.40	-1.29	0.12
1	6	1.27	2.01	0.38
1	7	0.33	0.47	0.13
1	8	0.43	-0.79	0.10
1	9	0.78	0.66	0.24
1	10	0.71	-0.15	0.14
1	11	0.28	2.67	0.20
1	12	1.60	1.83	0.14
1	13	0.83	1.80	0.22
1	14	1.66	1.68	0.23
1	15	1.10	-0.98	0.11
1	16	0.88	-0.79	0.08
1	17	1.08	1.96	0.31
1	18	0.26	0.57	0.09
1	19	1.28	1.19	0.25
1	20	0.38	0.07	0.07
1	21	0.82	0.79	0.29
1	22	1.91	2.02	0.18
1	23	0.58	2.18	0.16
1	24	0.56	-0.75	0.09
1	25	1.58	0.14	0.28
1	26	1.18	1.45	0.22
1	27	0.69	0.03	0.23
1	28	0.84	0.96	0.15
1	29	0.44	0.33	0.11
1	30	1.12	0.74	0.33
1	31	1.72	0.49	0.24
1	32	1.84	1.76	0.24
1	33	1.05	1.68	0.25

Form	**Item**	***a***	***b***	***c***
1	34	1.60	2.81	0.23
1	35	2.41	1.68	0.35
1	36	0.44	1.92	0.15
1	37	1.08	2.26	0.16
1	38	0.62	0.26	0.08
1	39	0.54	0.54	0.12
1	40	0.51	0.64	0.09
1	41	1.24	0.62	0.27
1	42	1.07	1.68	0.24
2	1	0.67	-0.30	0.24
2	2	0.49	0.04	0.24
2	3	0.55	0.78	0.29
2	4	0.28	-0.02	0.13
2	5	0.48	1.08	0.14
2	6	0.61	2.65	0.27
2	7	1.48	1.52	0.30
2	8	0.92	-0.09	0.21
2	9	0.79	-0.82	0.12
2	10	0.76	1.42	0.24
2	11	0.77	0.31	0.20
2	12	0.67	1.44	0.23
2	13	0.27	1.72	0.11
2	14	0.57	0.01	0.17
2	15	0.46	0.30	0.07
2	16	1.43	1.95	0.23
2	17	1.99	2.58	0.17
2	18	0.73	-0.64	0.14
2	19	0.53	2.40	0.19
2	20	0.99	0.43	0.24
2	21	0.86	0.69	0.25
2	22	0.38	1.18	0.11
2	23	1.52	2.25	0.16
2	24	0.60	1.68	0.13
2	25	1.36	0.93	0.19
2	26	0.89	-1.03	0.19
2	27	1.09	-1.25	0.08
2	28	1.09	-0.27	0.17
2	29	1.96	1.48	0.21
2	30	0.55	0.70	0.22
2	31	0.21	2.79	0.15
2	32	1.57	1.29	0.18
2	33	0.82	1.36	0.23
2	34	0.46	0.72	0.08
2	35	0.22	1.16	0.11
2	36	1.44	-0.12	0.28
2	37	0.37	2.02	0.12
2	38	1.37	-0.52	0.18

Form	Item	*a*	*b*	*c*
2	39	0.33	2.66	0.09
2	40	0.28	1.50	0.13
2	41	1.38	0.52	0.15
2	42	0.22	3.63	0.12
3	1	0.45	-1.47	0.18
3	2	1.46	0.57	0.42
3	3	1.07	3.55	0.33
3	4	1.50	2.26	0.10
3	5	0.80	-1.32	0.18
3	6	0.43	0.18	0.11
3	7	0.96	-0.50	0.27
3	8	1.27	1.32	0.24
3	9	1.06	-0.52	0.25
3	10	1.03	2.17	0.24
3	11	0.73	1.62	0.19
3	12	0.41	0.00	0.17
3	13	0.56	-0.97	0.07
3	14	1.50	-1.46	0.08
3	15	1.38	-0.07	0.28
3	17	0.58	-0.56	0.13
3	18	1.49	-0.64	0.32
3	19	1.46	0.80	0.22
3	20	1.29	0.80	0.28
3	21	1.06	-0.25	0.21
3	22	1.29	0.72	0.29
3	23	1.46	-1.29	0.10
3	24	1.74	-0.17	0.21
3	25	0.75	-1.11	0.11
3	26	0.83	0.58	0.15
3	27	0.87	-0.01	0.16
3	28	1.07	-0.16	0.18
3	29	1.10	0.23	0.21
3	30	0.59	1.10	0.32
3	31	1.45	1.05	0.23
3	32	0.87	-0.06	0.12
3	33	0.88	-1.14	0.08
3	34	0.67	-0.04	0.18
3	35	0.66	0.61	0.27
3	36	0.79	-1.13	0.13
3	37	1.67	-0.09	0.21
3	38	0.55	0.56	0.14
3	40	2.62	1.07	0.24
3	41	1.29	0.43	0.17
3	42	0.71	0.00	0.21
4	1	1.85	1.36	0.42
4	2	0.63	0.92	0.19
4	3	0.99	-0.01	0.30

Form	Item	*a*	*b*	*c*
4	4	2.00	2.45	0.12
4	5	0.79	1.12	0.27
4	6	0.82	-0.90	0.10
4	7	1.11	0.35	0.28
4	8	0.75	1.78	0.25
4	9	1.15	-0.16	0.20
4	10	0.54	4.76	0.25
4	12	0.95	-0.95	0.21
4	13	0.71	0.46	0.24
4	14	0.88	-0.49	0.07
4	15	0.89	0.77	0.15
4	16	1.03	-0.49	0.19
4	17	1.13	2.54	0.25
4	18	1.12	-0.10	0.19
4	19	1.25	0.07	0.18
4	20	0.83	-0.20	0.08
4	21	0.62	1.20	0.24
4	22	0.63	0.00	0.11
4	23	1.05	1.76	0.34
4	24	0.45	0.71	0.20
4	25	0.88	-0.34	0.14
4	26	0.90	1.04	0.19
4	27	0.59	3.99	0.15
4	28	0.66	2.70	0.27
4	29	0.88	-1.03	0.07
4	30	0.75	1.20	0.19
4	31	0.31	-0.21	0.09
4	32	0.52	0.37	0.07
4	33	0.19	3.06	0.10
4	34	0.75	0.48	0.24
4	35	0.58	0.88	0.14
4	36	0.63	0.31	0.10
4	37	1.41	0.40	0.26
4	38	0.90	0.61	0.13
4	39	0.23	1.72	0.12
4	40	1.17	-0.17	0.14
4	41	1.24	2.89	0.17
4	42	0.93	1.05	0.22
5	1	0.01	52.93	0.20
5	2	0.99	-0.59	0.42
5	3	0.25	-0.60	0.15
5	4	0.75	0.21	0.23
5	5	0.64	0.73	0.16
5	6	0.29	-0.70	0.09
5	7	0.68	2.64	0.31
5	8	0.95	0.37	0.26
5	9	1.52	0.20	0.24

Form	**Item**	***a***	***b***	***c***
5	10	0.84	0.82	0.23
5	11	0.85	0.64	0.23
5	12	0.97	1.94	0.34
5	13	0.79	-0.50	0.12
5	14	0.85	1.71	0.15
5	15	0.94	1.96	0.26
5	16	0.98	0.12	0.18
5	17	0.71	-0.88	0.06
5	18	0.77	-0.35	0.14
5	19	0.57	-0.02	0.10
5	20	0.56	-0.75	0.08
5	21	0.62	-0.20	0.12
5	22	0.75	-0.62	0.06
5	24	1.14	0.19	0.24
5	25	0.58	0.11	0.06
5	26	0.80	0.84	0.20
5	27	0.64	0.38	0.13
5	28	0.55	0.02	0.14
5	29	1.73	1.74	0.26
5	30	0.53	0.39	0.07
5	31	0.54	1.17	0.22
5	32	0.29	3.06	0.16
5	33	1.04	0.58	0.19
5	34	1.24	-0.19	0.18
5	35	1.11	0.08	0.28
5	37	1.12	0.08	0.31
5	38	1.16	1.09	0.26
5	39	1.05	0.58	0.19
5	40	0.42	1.36	0.15
5	41	0.58	0.84	0.19
5	42	1.24	0.20	0.26

Appendix 2. Bootstrap P-values for Form 1 Online and Paper Testing over 100 Replications

Item	P-value		Diff.	SD of Diff.	Standardized Diff.	Sig.	Effect Size
	Online	Paper					
1	0.75	0.77	-0.02	0.02	-0.70		-0.04
2	0.78	0.82	-0.04	0.02	-1.82		-0.09
3	0.36	0.40	-0.04	0.03	-1.53		-0.09
4	0.44	0.42	0.02	0.03	0.88		0.05
5	0.86	0.87	0.00	0.02	-0.29		-0.01
6	0.47	0.42	0.05	0.03	1.69		0.09
7	0.52	0.51	0.01	0.03	0.34		0.02
8	0.66	0.66	0.00	0.03	0.00		0.00
9	0.53	0.49	0.04	0.02	1.91		0.09
10	0.57	0.60	-0.03	0.03	-1.21		-0.06
11	0.42	0.37	0.05	0.03	1.55		0.09
12	0.18	0.19	-0.01	0.02	-0.39		-0.02
13	0.32	0.31	0.00	0.03	0.15		0.01
14	0.28	0.28	0.00	0.02	0.10		0.01
15	0.81	0.79	0.03	0.02	1.41		0.07
16	0.73	0.73	0.00	0.03	-0.04		0.00
17	0.37	0.36	0.02	0.03	0.62		0.03
18	0.49	0.50	-0.01	0.03	-0.30		-0.02
19	0.36	0.37	-0.01	0.03	-0.49		-0.03
20	0.57	0.53	0.04	0.03	1.33		0.07
21	0.47	0.50	-0.03	0.03	-0.88		-0.05
22	0.20	0.20	-0.01	0.02	-0.25		-0.01
23	0.25	0.28	-0.02	0.03	-0.92		-0.05
24	0.69	0.67	0.01	0.03	0.54		0.03
25	0.64	0.60	0.04	0.03	1.56		0.08
26	0.32	0.32	0.00	0.02	-0.06		0.00
27	0.64	0.61	0.04	0.02	1.47		0.07
28	0.31	0.37	-0.05	0.02	-2.33	*	-0.12
29	0.51	0.50	0.01	0.03	0.32		0.02
30	0.56	0.51	0.05	0.03	1.56		0.09
31	0.48	0.49	-0.01	0.03	-0.41		-0.02
32	0.29	0.28	0.01	0.03	0.36		0.02
33	0.27	0.32	-0.05	0.02	-2.04	*	-0.11
34	0.21	0.24	-0.02	0.02	-0.98		-0.06
35	0.42	0.39	0.04	0.02	1.67		0.08
36	0.34	0.33	0.01	0.03	0.56		0.03
37	0.19	0.19	-0.01	0.02	-0.30		-0.02
38	0.49	0.49	0.00	0.03	0.18		0.01
39	0.45	0.45	0.00	0.03	-0.09		0.00
40	0.43	0.43	0.00	0.03	-0.02		0.00
41	0.45	0.50	-0.05	0.02	-1.97	*	-0.10
42	0.31	0.32	-0.01	0.02	-0.32		-0.02

* The standardized difference is greater than or equal to 1.96.

Appendix 3. Bootstrap P-values for Form 2 Online and Paper Testing over 100 Replications

Item	p-value		Diff.	SD of Diff.	Standardized Diff.	Sig.	Effect Size
	Online	Paper					
1	0.65	0.67	-0.02	0.03	-0.77		-0.04
2	0.58	0.61	-0.03	0.03	-1.18		-0.06
3	0.49	0.54	-0.04	0.03	-1.47		-0.09
4	0.58	0.57	0.01	0.02	0.50		0.02
5	0.34	0.41	-0.06	0.03	-2.30	*	-0.13
6	0.29	0.33	-0.03	0.03	-1.29		-0.07
7	0.38	0.37	0.02	0.02	0.66		0.03
8	0.59	0.62	-0.03	0.02	-1.08		-0.05
9	0.72	0.73	-0.02	0.02	-0.78		-0.04
10	0.30	0.39	-0.09	0.02	-3.54	*	-0.18
11	0.55	0.53	0.01	0.03	0.46		0.02
12	0.42	0.38	0.04	0.03	1.41		0.08
13	0.40	0.40	0.00	0.03	-0.03		0.00
14	0.58	0.58	0.01	0.03	0.25		0.01
15	0.49	0.49	0.00	0.03	0.16		0.01
16	0.29	0.27	0.02	0.03	0.68		0.04
17	0.19	0.18	0.01	0.02	0.50		0.03
18	0.64	0.70	-0.06	0.03	-2.21	*	-0.12
19	0.30	0.29	0.01	0.03	0.20		0.01
20	0.49	0.52	-0.03	0.02	-1.32		-0.06
21	0.49	0.50	-0.01	0.02	-0.35		-0.02
22	0.40	0.40	0.00	0.03	0.03		0.00
23	0.19	0.18	0.02	0.02	0.75		0.04
24	0.32	0.29	0.03	0.02	1.32		0.06
25	0.34	0.37	-0.03	0.02	-1.30		-0.07
26	0.83	0.80	0.03	0.02	1.75		0.09
27	0.84	0.84	0.00	0.02	0.00		0.00
28	0.65	0.65	0.00	0.03	-0.04		0.00
29	0.28	0.28	0.00	0.02	-0.17		-0.01
30	0.43	0.51	-0.08	0.03	-2.60	*	-0.17
31	0.36	0.38	-0.02	0.03	-0.67		-0.04
32	0.30	0.29	0.01	0.03	0.31		0.02
33	0.43	0.38	0.05	0.03	1.85		0.10
34	0.41	0.43	-0.02	0.03	-0.69		-0.04
35	0.45	0.47	-0.02	0.03	-0.64		-0.04
36	0.68	0.67	0.01	0.02	0.65		0.03
37	0.35	0.32	0.02	0.03	0.83		0.05
38	0.76	0.72	0.04	0.02	2.06	*	0.10
39	0.28	0.27	0.01	0.02	0.40		0.02
40	0.41	0.42	-0.01	0.03	-0.19		-0.01
41	0.42	0.43	-0.01	0.03	-0.30		-0.02
42	0.24	0.29	-0.05	0.02	-2.09	*	-0.11

* The standardized difference is greater than or equal to 1.96.

Appendix 4. Bootstrap P-values for Form 3 Online and Paper Testing over 100 Replications

Item	p-value		Diff.	SD of Diff.	Standardized Diff.	Sig.	Effect Size
	Online	Paper					
1	0.79	0.78	0.02	0.02	0.73		0.04
2	0.56	0.61	-0.05	0.03	-1.66		-0.10
3	0.36	0.34	0.03	0.03	1.07		0.06
4	0.13	0.12	0.01	0.02	0.27		0.02
5	0.86	0.82	0.04	0.02	1.98	*	0.10
6	0.46	0.52	-0.06	0.03	-2.50	*	-0.13
7	0.71	0.72	-0.01	0.03	-0.41		-0.02
8	0.38	0.36	0.03	0.03	0.84		0.05
9	0.69	0.72	-0.04	0.02	-1.50		-0.08
10	0.30	0.28	0.02	0.02	0.84		0.05
11	0.28	0.32	-0.03	0.03	-1.23		-0.08
12	0.58	0.58	0.01	0.03	0.19		0.01
13	0.65	0.70	-0.05	0.02	-2.01	*	-0.11
14	0.91	0.88	0.02	0.02	1.38		0.07
15	0.66	0.65	0.01	0.02	0.56		0.03
16	0.22	0.21	0.01	0.02	0.61		0.04
17	0.69	0.66	0.03	0.02	1.46		0.07
18	0.76	0.79	-0.03	0.02	-1.45		-0.07
19	0.39	0.41	-0.02	0.02	-0.98		-0.05
20	0.44	0.47	-0.03	0.03	-1.09		-0.06
21	0.66	0.66	0.00	0.03	-0.10		-0.01
22	0.49	0.49	0.00	0.02	0.01		0.00
23	0.86	0.86	0.00	0.02	0.13		0.01
24	0.62	0.64	-0.03	0.02	-1.11		-0.05
25	0.75	0.76	-0.01	0.02	-0.43		-0.02
26	0.42	0.45	-0.02	0.03	-0.95		-0.05
27	0.60	0.56	0.04	0.03	1.53		0.08
28	0.59	0.62	-0.04	0.02	-1.47		-0.07
29	0.54	0.54	0.00	0.03	0.19		0.01
30	0.52	0.51	0.01	0.03	0.38		0.02
31	0.34	0.38	-0.04	0.03	-1.60		-0.09
32	0.55	0.56	-0.01	0.02	-0.43		-0.02
33	0.78	0.79	-0.01	0.02	-0.51		-0.03
34	0.64	0.60	0.04	0.03	1.60		0.09
35	0.55	0.54	0.01	0.02	0.41		0.02
36	0.82	0.79	0.03	0.02	1.42		0.08
37	0.57	0.62	-0.05	0.02	-1.93		-0.10
38	0.48	0.47	0.01	0.02	0.32		0.02
39	0.29	0.23	0.06	0.02	2.52	*	0.14
40	0.39	0.35	0.04	0.02	1.49		0.07
41	0.42	0.47	-0.05	0.02	-2.07	*	-0.10
42	0.58	0.60	-0.02	0.02	-0.87		-0.04

* The standardized difference is greater than or equal to 1.96.

Appendix 5. Bootstrap P-values for Form 4 Online and Paper Testing over 100 Replications

	p-value			**SD of**	**Standardized**		**Effect**
Item	**Online**	**Paper**	**Diff.**	**Diff.**	**Diff.**	**Sig.**	**Size**
1	0.50	0.49	0.01	0.03	0.21		0.01
2	0.46	0.44	0.02	0.03	0.69		0.04
3	0.61	0.65	-0.04	0.03	-1.37		-0.08
4	0.12	0.14	-0.01	0.02	-0.82		-0.04
5	0.38	0.44	-0.06	0.03	-1.87		-0.12
6	0.74	0.74	0.00	0.02	-0.09		0.00
7	0.56	0.57	-0.01	0.03	-0.46		-0.03
8	0.32	0.36	-0.04	0.02	-1.71		-0.08
9	0.67	0.64	0.03	0.02	1.04		0.05
10	0.29	0.25	0.04	0.02	1.78		0.09
11	0.17	0.14	0.04	0.02	1.80		0.10
12	0.81	0.80	0.01	0.02	0.51		0.03
13	0.52	0.53	-0.01	0.03	-0.45		-0.03
14	0.66	0.65	0.00	0.02	0.08		0.00
15	0.35	0.40	-0.05	0.03	-1.86		-0.10
16	0.74	0.71	0.03	0.02	1.20		0.06
17	0.28	0.27	0.00	0.02	0.20		0.01
18	0.56	0.62	-0.05	0.02	-2.33	*	-0.10
19	0.56	0.58	-0.02	0.02	-0.77		-0.03
20	0.59	0.59	0.00	0.02	0.00		0.00
21	0.45	0.45	0.00	0.03	0.09		0.01
22	0.56	0.56	0.00	0.02	-0.01		0.00
23	0.41	0.41	0.00	0.03	-0.18		-0.01
24	0.53	0.51	0.02	0.03	0.64		0.04
25	0.64	0.65	-0.01	0.02	-0.35		-0.02
26	0.45	0.39	0.06	0.03	2.14	*	0.13
27	0.15	0.18	-0.02	0.02	-1.00		-0.06
28	0.34	0.32	0.02	0.03	0.79		0.04
29	0.72	0.76	-0.04	0.02	-1.67		-0.08
30	0.41	0.38	0.03	0.03	1.16		0.07
31	0.56	0.56	0.00	0.03	0.11		0.01
32	0.45	0.46	-0.01	0.03	-0.37		-0.02
33	0.38	0.35	0.03	0.03	1.08		0.07
34	0.53	0.53	-0.01	0.03	-0.33		-0.02
35	0.36	0.42	-0.06	0.03	-2.11	*	-0.12
36	0.41	0.47	-0.06	0.03	-2.46	*	-0.13
37	0.50	0.54	-0.04	0.03	-1.61		-0.08
38	0.41	0.43	-0.01	0.02	-0.45		-0.02
39	0.46	0.42	0.04	0.03	1.42		0.08
40	0.64	0.62	0.02	0.02	0.95		0.05
41	0.19	0.18	0.02	0.02	0.67		0.04
42	0.43	0.41	0.02	0.03	0.76		0.04

* The standardized difference is greater than or equal to 1.96.

Appendix 6. Bootstrap P-values for Form 5 Online and Paper Testing over 100 Replications

Item	p-value		Diff.	SD of Diff.	Standardized Diff.	Sig.	Effect Size
	Online	Paper					
1	0.44	0.38	0.06	0.03	2.20	*	0.13
2	0.82	0.79	0.03	0.02	1.50		0.08
3	0.63	0.62	0.01	0.03	0.37		0.02
4	0.48	0.58	-0.09	0.03	-3.48	*	-0.19
5	0.41	0.45	-0.03	0.03	-1.29		-0.07
6	0.60	0.62	-0.02	0.03	-0.75		-0.05
7	0.35	0.36	-0.01	0.03	-0.25		-0.02
8	0.56	0.56	0.01	0.03	0.28		0.02
9	0.57	0.58	-0.01	0.02	-0.56		-0.03
10	0.45	0.47	-0.01	0.03	-0.52		-0.03
11	0.52	0.50	0.02	0.03	0.64		0.04
12	0.49	0.39	0.10	0.03	3.63	*	0.20
13	0.68	0.67	0.01	0.02	0.57		0.03
14	0.33	0.27	0.06	0.02	2.58	*	0.14
15	0.35	0.34	0.02	0.03	0.54		0.03
16	0.58	0.57	0.01	0.03	0.31		0.02
17	0.77	0.71	0.06	0.02	2.75	*	0.15
18	0.61	0.64	-0.04	0.03	-1.25		-0.08
19	0.60	0.56	0.04	0.03	1.27		0.07
20	0.71	0.67	0.04	0.03	1.37		0.08
21	0.62	0.59	0.02	0.03	0.92		0.05
22	0.68	0.66	0.02	0.03	0.78		0.04
23	0.15	0.15	0.00	0.02	0.04		0.00
24	0.57	0.59	-0.02	0.03	-0.66		-0.04
25	0.52	0.51	0.01	0.03	0.28		0.02
26	0.47	0.43	0.03	0.03	1.22		0.06
27	0.54	0.49	0.05	0.03	2.10	*	0.11
28	0.59	0.57	0.02	0.03	0.81		0.04
29	0.32	0.32	0.00	0.03	0.18		0.01
30	0.49	0.47	0.02	0.03	0.57		0.03
31	0.49	0.44	0.05	0.03	1.57		0.09
32	0.36	0.32	0.04	0.03	1.35		0.07
33	0.44	0.47	-0.02	0.03	-0.80		-0.04
34	0.65	0.64	0.01	0.03	0.54		0.03
35	0.66	0.62	0.03	0.03	1.19		0.07
36	0.18	0.18	0.00	0.02	0.18		0.01
37	0.68	0.65	0.03	0.02	1.47		0.07
38	0.38	0.41	-0.03	0.03	-0.92		-0.05
39	0.50	0.47	0.03	0.03	1.13		0.06
40	0.42	0.40	0.02	0.03	0.62		0.04
41	0.48	0.46	0.02	0.03	0.69		0.04
42	0.61	0.58	0.02	0.03	0.71		0.04

* The standardized difference is greater than or equal to 1.96.