Mathematics

Minnesota Comprehensive Assessment-Series III (MCA-III)

Mode Comparability Study Report

February 15, 2012





Introduction

When testing programs use scores obtained through different modes of administration, such as online and paper-and-pencil, it is essential that a comparability study be conducted to ascertain how testing mode affects student performance. Administrations of Mathematics Minnesota Comprehensive Assessments Series III (MCA-III) are available in online and paper versions. It is therefore important to evaluate potential mode effects between the two versions of Mathematics MCA-III. This document reports on the Mathematics MCA-III comparability study conducted during the spring and summer of 2011.

Background

Whenever paper-based and online assessments co-exist, professional testing standards indicate the need to ensure comparable results across paper and online mediums. The Guidelines for Computer-Based Tests and Interpretations (APA, 1986) states: "...when interpreting scores from the computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cut scores obtained from conventional tests." (p. 18). The joint Standards for Educational and Psychological Testing also recommend empirical validation of score interpretations across computer-based and paper-based tests (AERA, APA, NCME, 1999, Standard 4.10).

Virtually all studies assessing the comparability of online and paper assessments utilize one of three general designs: (a) randomly equivalent groups; (b) test-retest; or (c) matched groups. Each of the three designs has different strengths and weaknesses which make them more or less desirable in specific circumstances. Features of the three designs are given in Figure 1 and described in the following paragraphs.

Figure 1. Features of Comparability Design Options.

Design	Features	Potential Disadvantages
Randomized Groups	 Students are randomly assigned to either the paper or computer version. 	 Random assignment might be intrusive to districts and schools.
	 With well designed study, robust inferences can be drawn. 	
Test – Retest	 Each student in study takes both computer and paper version. Motivation is increased if student is awarded higher of two scores. In strongest version of design, two factors are counterbalanced—order of administration and test form. This means four separate groups are required: Computer (Form 1) – Paper (Form 2) Paper (Form 1) – Computer (Form 2) Computer (Form 2) – Paper (Form 1) Paper (Form 2) – Computer (Form 1) 	 Requires two test forms to be developed/exposed. Design becomes much weaker if counterbalancing cannot be achieved, especially if computer and paper version are different forms. Extra testing is burdensome to schools/students. Susceptible to fatigue and motivation effects, especially w/o counterbalancing.
Matched Groups	 Quasi-experimental design where no random assignment is done for the two groups. Comparison of groups is accomplished by matching groups on external variable, such as a previous test score. Pearson has found that correlations of .78 between matching variable and student performance are sufficient. 	 Inferences drawn are reasonable only to the degree that matching variable is effective. Does not control for other confounding differences between the groups.

In the randomized groups design, students are randomly assigned to test in either online or paper testing groups. When this design is feasible (and sample sizes are sufficiently large), it is the strongest of the three alternate designs. However, this design is intrusive to districts and schools, and the researchers typically must exert a high degree of control to ensure that all participating students are randomly assigned to the online or paper condition.

In the test-retest comparability study design, participating students test twice within a short period of time, once with a test form administered online and once with an alternate paper test form. The advantage of this design is that students are typically offered the higher of the two scores they obtain, which ensures that they are not disadvantaged by testing online, even if the online tests result in lower scores on average. In the strongest version of this design, both the test forms and the order of administration are counterbalanced. However, it is sometimes not

feasible to counterbalance the test forms, and a more commonly used and much weaker version of this design is to administer one form in paper format (e.g., the operational form) and an alternate form online. In addition, it is not always possible to counterbalance order of administration within a school, which further weakens the design. Finally, schools and students are often reluctant to accept the additional burden of two different administrations of the same test, and those that do participate are often affected by fatigue or motivation, resulting in mode by sequence interaction effects.

The matched groups design is a quasi-experimental design in which meaningful comparisons between the online and paper group are made possible by matching the groups on one or more external variables, such as a previous test score. In this design, the same test form is 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 who take the paper 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 successfully employed the matched groups design using scores from the previous test as the matching variable. Pearson has found that correlations between scores in consecutive years typically run between 0.7 and 0.8, and that this relationship is strong enough to make prior year score an effective covariate for comparing the online and paper groups (Way, Davis, & Fitzpatrick, 2006). However, additional demographic variables, such as gender and ethnicity, could be included for purposes of matching groups taking online and paper forms.

Methodology

Students taking the 2011 Mathematics MCA-III took the test in either the online or paper format. The choice was made at the district level and thus testing mode was not randomly assigned. Because groups were not randomly assigned, a matched groups design was implemented. One of the 20 online forms in each grade had nearly complete item overlap with the paper form and was used to compare online and paper modes of administration.

The matched samples comparability analyses (MSCA; Way et al., 2006) design was used as the basis for examining the comparability of the online and paper forms. Pearson has successfully used the MSCA matching strategy for conducting comparability studies in other contexts, and the flexibility of the MSCA approach, which can be applied to a variety of data and statistical analyses, is a particular advantage for examining comparability of paper and online testing modes for Mathematics MCA-III.

MSCA is a bootstrapping approach that creates matched samples and derives estimates of the random error due to sampling. Students' previous assessment scores and/or demographic variables are used as matching variables to obtain sub-samples of students from one mode that equal the numbers of students testing under the mode with fewer students. For example, if 2,000 students completed the online form and 5,000 students completed the paper form, a bootstrap sample of 2,000 students would be drawn from the paper form to match a bootstrap sample of students taking the online form. Students testing under one mode are selected so that

matching variable characteristics are as similar as possible to those testing under the other mode. A regression equation is developed using matching variables as predictors of test score. Then students in one group are matched to students in the other group using the predicted score on that weighted set of predictors.

In this study, a version of MSCA was implemented using the bootstrapping approach where a sub-sample of students in the paper group were matched to students from the online group using their previous mathematics and concurrent reading test scores, gender, ethnicity, and free/reduced price lunch status as matching variables. The comparability study involved two stages. The first stage evaluated mode comparability in each grade and the second stage implemented a plan for equating tests in cases where mode differences were found. Details of the methodology are outlined below.

Stage 1: Comparability between online and paper using matched samples of students

Setting up the matching variable:

 For all eligible students1 who tested in the paper mode, their 2011 Mathematics MCA-III raw scores2 were regressed on their mathematics and/or reading scale scores from the previous (mathematics) and current test administration (reading). In addition, demographic information, including free and reduced price lunch status (FRL) and ethnic group membership were included in the regression equation to control for possible differences in student characteristics across those students taking online versus paper forms.

For students in grade 4 – 8, who have prior MCA-II Mathematics scores, the regression equation is

$$\hat{Y}_{predicted_rawscore} = b_0 + b_1 X_{1(2011MCA-II_Reading)} + b_2 X_{2(2010MCA-II_Math)} + b_3 X_{3(Am.Indian-White)} + b_4 X_{4(Asian-White)} + b_5 X_{5(Hispanic-White)} + b_6 X_{6(Black-White)} + b_7 X_{7(Female-Male)} + b_8 X_{8(FRL-NonFRL)} + b_8 X_{8(FRL-NoF$$

where X_i s are values on the matching variable predictors, b's are estimated regression weights, and y-hat is the predicted Mathematics MCA-III raw score.

¹ Students taking the paper version of the 2011 Mathematics MCA-III test with accommodations were not included in the samples used for the comparability study. Overall, accommodated student performance tends to be lower than that for non-accommodated students. Because accommodated students are disproportionally assigned to the paper mode, including accommodated students would have created greater disparity in proficiency between the paper and online groups, which could have resulted in difficulties in creating matched samples. For this reason, accommodated students were excluded from the study.

² Students with missing values on any of the matching variables were not included in the study. In addition, the 2011 Mathematics MCA-III raw score was computed using only items common between modes.

For students in grade 3, who do not have prior MCA-II Mathematics scores, the regression equation is

 $\hat{Y}_{predicted_rawscore} = b_0 + b_1 X_{1(2011MCA-II_Reading)} + b_2 X_{2(Am.Indian-White)} + b_3 X_{3(Asian-White)} + b_4 X_{4(Hispanic-White)} + b_5 X_{5(Black-White)} + b_6 X_{6(Female-Male)} + b_7 X_{7(FRL-NonFRL)}$

- 2. In each grade, the regression coefficients were estimated and applied to all eligible students (paper and online) to obtain a predicted raw score (y-hat) for each student. Using this approach, the online students' predicted scores were generated using regression weights obtained from students taking the paper test. The regression coefficients and multiple correlation coefficient for each grade are reported in Appendix A. Across the grades, the estimated R2 ranged from .56 to .72, and thus were of a magnitude that Pearson had found adequate in previous MSCA studies.
- 3. Students were then broken into 20 groups based on the predicted raw score (y-hat). Score group categories were defined by dividing the online predicted raw score distribution into 20 equal sized groups. Paper students were then divided into groups using the score categories defined by the online distribution. The grouping procedure was used as the basis for creating matched samples as part of the MSCA process described next.

The MSCA Procedure:

- 1. Using the group membership developed from the regression equation (step 3 above), a sample of students is drawn at random with replacement within each score group from the online participants.3
- 2. A matched sample is drawn at random with replacement from the available paper participants so that the sample size for each paper group equals the sample size for the corresponding online group.
- 3. To evaluate the adequacy of matching, matched samples are compared using the following descriptive statistics:
 - a. Summary statistics (e.g., means, standard deviations) of scores on the common items between online and paper.
 - b. Item level statistics (e.g., p-value for each of common item).
- 4. Test performance and scaling results are compared across matched samples:
 - a. Mean raw score differences (see Appendix B).
 - b. Effect sizes. (see Appendix B)

³ This sample is a bootstrap sample; that is, a sample drawn with replacement. The size of this sample is equal to the total size of the group.

- c. The item response theory (IRT) Stocking-Lord scale transformation (Stocking & Lord, 1983) between the matched paper samples and the full paper sample group, each calibrated separately using the 3-parameter IRT model. (Further explanation of this comparison, and how it would be used across replications, is described under point 6 below).
- 5. Steps 1 to 4 are repeated until the desired number of replications has been reached. In this study, 100 replications were done.
- 6. The statistics saved in step 4) are summarized across replications.
 - a. For 4a to 4b above, the replications create a bootstrap sampling distribution for matched online and paper groups that allows the use of statistical significance tests, such as the z-statistic (see Appendix).
 - b. For 4c above, a bootstrap sampling distribution is created using the Stocking-Lord scale transformations of the matched paper samples to the full paper sample. In scaling matched paper samples to the full paper sample, there can be no mode effect, but at the same time matched paper samples are representative of proficiency levels in the online samples. Thus, this sampling distribution gives the expected spread of scale transformations if one were equating the online form to the full-sample paper form metric, in the absence of any mode effect. To evaluate a potential mode effect, the full sample online to full sample paper scale transformation is estimated and compared to the scale transformation values in the sampling distribution. A mode effect would be suggested if the observed online to paper transformation was found to be an outlier in relation to the sampling distribution.

Stage 2: Calibration and scaling of item parameters between online and paper groups

In this stage any mode differences detected in Stage 1 are accounted for in the scaling of the paper administration to the online test. This is accomplished through the use of a common item equating design. The basic approach is to determine a set of items in common between the online and paper forms that do not indicate a mode effect. These items are used to link the paper form to the online form.

To determine which items in common between the online and paper forms did not indicate a mode effect, a number of factors were considered. The primary factor was the item's delta difference between matched paper and online samples. Delta is found by converting the proportion incorrect (1-p) to a z-score and using the following formula,

$$Delta = 13 + 4 \times z$$

Delta is an inverse normal transformation of the percent correct to a linear scale with a mean of 13 and a standard deviation of 4.

Differences in online and paper delta values for an item were computed. Delta differences greater than 0.4 were considered suggestive of a possible mode effect. However, items

displaying small delta differences still might be excluded from the linking set for other reasons, such as if the sequence positions between the online and paper versions greatly differed or if the item appeared in disparate sequence positions in the online forms in which it appeared.

Calibration for the two groups (paper and online) went as follows. First, the item parameters for online and paper forms were estimated separately. Stocking-Lord equating was then performed, using the identified linking set. The online group was considered to be the base group, which meant the online test determined the IRT scale. The Stocking-Lord transformation constants were then applied to the paper-form calibration item parameter estimates for items not included in the linking set. Online-form item parameters were assigned to the paper form items that had been included in the linking set.

For all analyses in Stage 1 and Stage 2, two Pearson data analysts programmed independent versions of the software to conduct the analysis as a matter of quality assurance. This seemed an especially important precaution to take given the complicated programming required for the analyses.

Results

Table 1 shows the test level analysis using the mean mathematics raw score of matched samples of online and paper groups across 100 iterations. The absolute mean mathematics raw score differences ranged from 0.18 in grade 8 to 1.06 in grade 5. Across all grades, the students who took paper version of the test tended to score higher on the common items than those who took the online version of the test, with students scoring on average about one raw score point higher on the paper versions in grades 3-6 and less than a raw score point in grades 7 and 8. There is a significant mean raw score difference between online and paper modes in grades 3 through 7. A similar trend of differences between modes can be observed in the effect sizes, which are based on the average effect size over 100 replications within each grade. The negative values of effect sizes indicate that paper samples performed higher on average than the online samples. However, the magnitudes of effect sizes were relatively small, using the criteria of 0.3 as a small effect (Cohen, 1988).

Grade	Matched Online Math Mean	Matched Paper Math Mean	Difference Math Mean	Difference Math SD	z Statistic	Flag*	Effect Size
3	34.74	35.71	-0.97	0.17	-5.71	*	-0.12
4	30.96	31.92	-0.96	0.15	-6.22	*	-0.13
5	32.45	33.51	-1.06	0.16	-6.61	*	-0.12
6	30.94	31.98	-1.04	0.14	-7.55	*	-0.12
7	28.87	29.46	-0.59	0.17	-3.50	*	-0.06
8	28.76	28.94	-0.18	0.13	-1.45		-0.02

 Table 1. Matched sample math mean raw score across 100 iterations

* |Z statistic| ≥ 2

Two 3PL IRT scaling analyses were conducted. In one analysis, a Stocking-Lord (SL; Stocking & Lord, 1983) scaling was carried out for each replication that scaled (equated) the matched bootstrap paper sample metric to the full paper sample metric. This analysis served as the baseline condition where the equating adjusted for ability distribution differences in the absence of a mode effect. After averaging over the 100 replications, if the SL slope and intercept constants differed significantly from the identity function (slope=1, intercept=0), this would indicate that the full paper and matched-sample paper ability distributions differed (and imply that the full paper and online samples also differed). In the second analysis, the unmatched online sample metric was equated to the full paper sample metric. In this analysis, both mode and ability distribution differences will impact the scaling. Therefore, comparing the observed scaling constants to those from the baseline matched paper to full paper equating condition in the first analysis provides an indication of mode effect.

Table 2 presents the average of the Stocking-Lord slopes and intercepts across the 100 replications of the matched bootstrap paper to full paper scaling, as well as the results of the online to paper scaling. Bootstrap standard errors (standard deviations of the scaling constants across replications) are also provided. The matched bootstrap paper sample constants can be compared to the identity function, which would be the expected result if the paper and online ability distributions were identical. In general, the matched paper slope values are within two standard errors of the identity function, indicating that the standard deviations of the ability distributions do not meaningfully differ. Note, however, the intercept values are within two standard errors for grades 3 and 4, but not for grades 5-8. These results indicate that the matched paper distribution (and by extension, the online distribution) is of slightly higher ability than the full paper sample for grades 5 to 8. Using Table 2 to compare the online and matched paper SL scaling coefficients, it can be seen that for all grades the online slope constants are within two standard errors of the matched paper slope constants, indicating that any mode effect doesn't have a significant impact on slope. For the intercept constants, however, none of the online values are within two standard errors of the matched paper values. The values especially differ in the lower grades, where the results indicate that online students were negatively impacted by mode effects on the order of 0.15 theta (or paper sample SD) units.

Grade	Sample	Slope	Intercept	SE of slope	SE of intercept
3	Matched Paper	1.01	-0.01	0.016	0.016
5	Online	0.99	-0.17		•
1	Matched Paper	1.00	0.02	0.016	0.016
4	Online	0.97	-0.12		
5	Matched Paper	0.97	0.05	0.017	0.015
5	Online	0.96	-0.08		
6	Matched Paper	1.03	0.07	0.015	0.013
0	Online	1.05	-0.06		
7	Matched Paper	0.99	0.04	0.015	0.013
1	Online	1.00	-0.04		
0	Matched Papers	1.01	0.04	0.015	0.011
0	Online	1.02	0.01		

Table 2. Stocking-Lord slope and intercept constants comparisons: Average of bootstrapmatched paper samples vs. full paper sample, and online sample vs. full paper sample*

* Paper form is used as the base form.

Another perspective of the IRT scaling results is given in Figure 2 through Figure 7. Instead of averaging over replications as in Table 2, the figures display all 100 results for the matched bootstrap paper to full paper scalings (these are represented by blue circles with an "MP" label). In addition, the identity function is shown as an "X" with the label "Paper" and the online to paper scaling is demarcated with the symbol "+" and the label "Online". Having the "Paper" point within the swarm of blue points would indicate that the proficiency distributions of students taking the two modes were similar. This is the case for grades 3 and 4, but not for the higher grades. Having the "Online" point within the swarm of blue points would be indicative that there was no significant mode effect. However, on the intercept axis, the "Online" point is clearly not in the swarm for any of the grades; the difference is the smallest for grade 8.





Figure 3.



Figure 4.



Figure 5.



Figure 6.



Figure 7.



The results of the Stage 1 comparability analyses given in Table 1, Table 2, and Figures 2-7 provide evidence that a mode effect would have had the potential to effect online scores in most, if not all, grades, although effect sizes might be considered to be relatively small. Given these empirical results, it was determined that the Stage 2 calibration and scaling of item parameters would be appropriate for all grades. This would allow any potential mode effect to be accounted for, making online and paper scores more comparable and thereby enhancing fairness across modes of administration.

To place the online and paper item calibrations on the same scale, a linking set was created in each grade, composed of items that were determined not to be impacted by a mode effect. Using the MSCA methodology, item level results were obtained comparing the results of the matched bootstrap paper samples with the bootstrap online samples. Item delta differences greater than 0.4 between online and paper groups were considered suggestive of a mode effect, and led to exclusion of items from the linking set. As described previously, a few additional items were excluded from the linking set if the sequence positions between the online and paper versions greatly differed or if the item was in disparate sequence positions in the online forms in which it appeared. The number of common items selected for the linking set ranged from 23 to 30 items across grades.

Items from the online and paper administrations were separately calibrated. Using items from the linking set, the paper test was scaled to the online test using Stocking-Lord. Because the items in the linking set were not impacted by the mode effect, this scaling accounted only for the proficiency differences between the online and paper samples and not for the mode effect. Results are presented in Table 3. The scaling constants are close to the identify function (1,0), indicating that the two groups did not differ much in ability 4 .

Grade	Slope	Intercept
3	1.021716	0.026471
4	1.013704	0.027288
5	1.007468	0.021827
6	0.962933	-0.047593
7	0.973433	0.007149
8	1.027296	-0.024554

Table 3. Stocking-Lord slope and intercept constants scaling paper test to online test using linking set composed of mode-neutral items

Because items in the linking set were judged to not be affected by mode, the online item parameter estimates were used as the final banked parameter estimates for the paper version of these items as well. Using the online parameters for the paper versions would not be appropriate for items not in the linking set, however, as these items were impacted by mode. Therefore, for items not in the linking set, the estimates from the separate paper item calibration were used, *after* transforming the parameters to the online scale using the values in Table 3. Final student scores were computed using either the online or the paper set of item parameters, as appropriate, for the mode taken by a given student.

Conclusion

This study was conducted to examine the comparability of scores from paper and online administrations of the Mathematics MCA-III. Although the results indicated the presence of relatively small overall mode effects that favored the paper administration, these effects were observed for a minority of items common to the paper and online forms. The approach used to adjust for these mode effects was essentially to treat the online and paper versions of the affected items as distinct items, with mode-specific parameter estimates. By using a set of linking items not impacted by mode differences, the paper mode-specific and/or unique items could be scaled to the online scale to account for population differences between online and paper groups.

The 2011 Mathematics MCA-III administration was unique in that essentially identical paper and fixed-form online tests were used operationally, permitting the analytical approaches used in the

⁴ The paper test was calibrated after excluding students who took accommodated forms so as to make the paper and online ability distributions more similar.

study. Beginning in 2012, the fixed-form paper format will continue in use, while the online format test will be computer-adaptive. Going forward, our expectation is that participation on paper forms will show the rapid decline seen in other states when online adaptive tests are deployed. Building on the steps outlined above, which were taken to enhance the comparability of paper and online tests in 2011, we plan to maintain comparability of paper and on-line scores by equating subsequent paper forms to the baseline 2011 paper form.

Appendix A: Model Fit and Regression Coefficients

Model Fit	Estimate (Sig)	F-statistic	P-value
R ²	0.5602*	2252.89	< 0.0001
Adj. R ²	0.5599*		

Table A1. Grade 3 model fit

* significant at 0.05 α -level.

Table A2. Grade 3 regression coefficients

Variable	Estimate	SE	t-statistic	P-value
Intercept	12.5792*	0.2624	47.93	< 0.0001
2011 MCA-II Reading	0.7126*	0.0068	104.69	< 0.0001
Am. Indian – White	-0.5044	0.3972	-1.27	0.2041
Asian – White	0.0151	0.1790	0.08	0.9328
Hispanic – White	-0.8206*	0.1864	-4.40	< 0.0001
Black – White	-2.1792*	0.1610	-13.54	< 0.0001
Female – Male	-1.6317*	0.0920	-17.74	< 0.0001
FRP – Non-FRP	-1.1494*	0.1096	-10.48	< 0.0001

* significant at 0.05 α -level

Table A3. Grade 4 model fit

Model Fit	Estimate (Sig)	F-statistic	P-value
R ²	0.6520*	2690.80	< 0.0001
Adj. R ²	0.6517*		

* significant at 0.05 α -level.

Table A4. Grade 4 regression coefficients

Variable	Estimate	SE	t-statistic	P-value
Intercept	0.7789*	0.2718	2.87	0.0042
2010 MCA-II Math	0.5229*	0.0076	68.62	<0.0001
2011 MCA-II Reading	0.3475*	0.0087	39.91	< 0.0001
Am. Indian – White	0.2583	0.3863	0.67	0.5037
Asian – White	1.1934*	0.1702	7.01	<0.0001
Hispanic – White	0.0887	0.1919	0.46	0.6439
Black – White	-0.1957	0.1497	-1.31	0.1912
Female – Male	-0.3657*	0.0852	-4.29	< 0.0001
FRP – Non-FRP	-0.6757*	0.0998	-6.77	< 0.0001

* significant at 0.05 α -level

Table A5. Grade 5 model fit

Model Fit	Estimate (Sig)	F-statistic	P-value
R ²	0.6861*	3042.76	< 0.0001
Adj. R ²	0.6859*		

* significant at 0.05 α -level.

Table A6. Grade 5 regression coefficients

Variable	Estimate	SE	t-statistic	P-value
Intercept	-1.0828*	0.2888	-3.75	0.0002
2010 MCA-II Math	0.5844*	0.0078	75.08	< 0.0001
2011 MCA-II Reading	0.3627*	0.0087	41.83	< 0.0001
Am. Indian – White	-0.3014	0.4324	-0.70	0.4857
Asian – White	0.9910*	0.1788	5.54	< 0.0001
Hispanic – White	-0.7117*	0.2122	-3.35	0.0008
Black – White	-1.2139*	0.1651	-7.35	< 0.0001
Female – Male	-0.5408*	0.0930	-5.82	< 0.0001
FRP – Non-FRP	-0.6710*	0.1110	-6.05	< 0.0001

* significant at 0.05 α -level

Table A7. Grade 6 model fit

Model Fit	Estimate (Sig)	F-statistic	P-value
R ²	0.6803*	2960.63	< 0.0001
Adj. R ²	0.6801*		

* significant at 0.05 α -level.

Table A8. Grade 6 regression coefficients

Variable	Estimate	SE	t-statistic	P-value
Intercept	-2.5953*	0.2807	-9.25	< 0.0001
2010 MCA-II Math	0.5554*	0.0069	80.81	< 0.0001
2011 MCA-II Reading	0.2916*	0.0073	40.12	< 0.0001
Am. Indian – White	-1.2489*	0.3790	-3.30	0.0010
Asian – White	0.5607*	0.1785	3.14	0.0017
Hispanic – White	-0.7305*	0.2087	-3.50	0.0005
Black – White	-0.9451*	0.1622	-5.83	< 0.0001
Female – Male	-0.4153*	0.0901	-4.61	< 0.0001
FRP – Non-FRP	-0.3984*	0.1072	-3.72	0.0002

* significant at 0.05 α -level

Table A9. Grade 7 model fit

Model Fit	Estimate (Sig)	F-statistic	P-value
R ²	0.7157*	3319.98	< 0.0001
Adj. R ²	0.7155*		

* significant at 0.05 α -level.

Table A10. Grade 7 regression coefficients

Variable	Estimate	SE	t-statistic	P-value
Intercept	-4.8523*	0.2808	-17.28	< 0.0001
2010 MCA-II Math	0.5664*	0.0065	87.25	< 0.0001
2011 MCA-II Reading	0.2730*	0.0078	35.37	< 0.0001
Am. Indian – White	-0.7662	0.4113	-1.86	0.0625
Asian – White	0.1538	0.1912	0.80	0.4213
Hispanic – White	-0.5752*	0.2258	-2.55	0.0108
Black – White	-0.7722*	0.1811	-4.26	< 0.0001
Female – Male	-0.9557*	0.0968	-9.87	< 0.0001
FRP – Non-FRP	-0.5726*	0.1160	-4.93	< 0.0001

* significant at 0.05 α -level

Table A11. Grade 8 model fit

Model Fit	Estimate (Sig)	F-statistic	P-value
R ²	0.7022*	3220.95	< 0.0001
Adj. R ²	0.7020*		

* significant at 0.05 α -level.

Table A12. Grade 8 regression coefficients

Variable	Estimate	SE	t-statistic	P-value
Intercept	-0.9816*	0.2507	-3.91	< 0.0001
2010 MCA-II Math	0.5533*	0.0061	91.00	< 0.0001
2011 MCA-II Reading	0.2061*	0.0068	30.36	< 0.0001
Am. Indian – White	-1.2768*	0.3776	-3.38	0.0007
Asian – White	-0.2316	0.1750	-1.32	0.1857
Hispanic – White	-0.9812*	0.2216	-4.43	< 0.0001
Black – White	-0.6491*	0.1691	-3.84	0.0001
Female – Male	-0.3480*	0.0911	-3.82	0.0001
FRP – Non-FRP	-0.4067*	0.1107	-3.67	0.0002

* significant at 0.05 α -level

Appendix B: Computation of Mean Differences and Effect Sizes

Examinee differences between online and paper groups were examined at the total test score level and at the item level. Total test score level analysis compared differences in summary statistics, such as mean scores, between online and paper matched samples across replications. Item-level analysis compared differences in item statistics between online and paper matched samples across replications. The two statistics summarized below were used for comparing both score-level and item-level differences between testing modes.

z-statistic

The z-statistic was used to determine the statistical reliability of differences found between online and paper groups for both total test score level and item level analyses. The general formula for the z-statistic is given as

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

where \overline{D}_{Diff} is the grand mean of the difference being calculated between online and paper groups over replications; and SE_{diff} is the bootstrap standard error of mean differences over replications, which is simply the standard deviation of the difference scores over replications.

Effect size

The effect size between for the difference between means is calculated by the following equation:

$$Effectsize = \frac{\overline{X}_{Online} - \overline{X}_{Paper}}{\sqrt{\frac{\left(SD_{Online}^{2} + SD_{Paper}^{2}\right)}{2}}}$$

where \overline{X} is the mean of the variable of interest and *SD* is the standard deviation.

References

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

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (second ed.). <u>Lawrence Erlbaum Associates</u>.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201 - 210.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). *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.