Graduation-Required Assessment for Diploma (GRAD) Mathematics Comparability Study Report

October 28, 2009

Minnesota Department of Education

Graduation-Required Assessment for Diploma (GRAD) Mathematics Comparability Study Report

When testing programs use scores obtained through different modes of administration, it is appropriate to conduct a comparability study in order to evaluate how the testing mode affects student performance. The initial administration of the Graduation-Required Assessment for Diploma (GRAD) is embedded in the paper-based Minnesota Comprehensive Assessments-Series II (MCA-II) test, whereas subsequent stand-alone GRAD retests are administered online. It is important to determine if a mode difference between the two versions of the GRAD exists that would justify mitigation through statistical adjustment during equating of the online retests. This report summarizes results from the GRAD Mathematics comparability study conducted in May 2009 and follows the same general format as the GRAD Reading comparability report.

Background

Whenever paper-based and online assessments coexist, professional testing standards indicate the need to ensure comparable results across paper and online mediums. The *Guidelines for Computer-Based Tests and Interpretations* of the American Psychological Association (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 recommends 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; and (c) matched groups. Each of these three designs has different strengths and weaknesses, making them more or less desirable in specific circumstances. Features of the three designs are given in Figure 1 and are described in the following paragraphs.

Design	Features	Potential Disadvantages
Randomized Groups	 Students are randomly assigned to either the paper or computer version. With a well-designed study, robust inferences can be drawn. 	• Random assignment might be intrusive to districts and schools.
Test-Retest	 Each student in the study takes the computer and paper version. Motivation is increased if the student is awarded the higher of two scores. In the strongest version of the 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. The design becomes much weaker if counterbalancing cannot be achieved, especially if the computer and paper versions are different forms. Extra testing is burdensome to schools/students. Susceptible to fatigue and motivation effects, especially without 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 an external variable, such as a previous test score. Pearson has found that correlations of 0.7–0.8 between the matching variable and student performance are sufficient. 	 Inferences drawn are reasonable only to the degree that the matching variable is effective. Does not control for other potential confounding differences between the groups.

Figure 1. Features of Comparability Design Options

In the randomized groups design, students are randomly assigned to test either online or by paper and pencil. 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-andpencil 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-and-pencil test form. The advantage of this design is that students are typically offered the higher of the two scores they obtain, ensuring 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, 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-and-pencil format (e.g., the operational form) and an alternate form online. In addition, it is not always possible to counterbalance the order of administration within a school, further weakening the design. Finally, schools and students are often reluctant to accept the additional burden of two different administrations of the same test, and those who do participate are often affected by fatigue or motivation, resulting in mode by sequence interaction effects.

The matched groups design is really a quasi-experimental design in which meaningful comparisons between the online and paper-and-pencil groups are made possible by matching the groups on an external variable, such as a previous test score. In this design, the same test form is typically administered to the online and paper-and-pencil 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 subsample of the students who take the regular paper-and-pencil 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.

The state's testing contractor, Pearson, has successfully employed the matched groups design using scores from the previous spring's 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-and-pencil groups. Pearson has implemented this method in a streamlined fashion that permits the adjustment of the equating conversion table for the online group should mode differences be detected (Way, Davis, and Fitzpatrick, 2006).

Study Design

The randomized groups design was chosen for the GRAD comparability study. This is the strongest type of comparability design when conditions permit random assignment of students. MDE determined that school districts in Minnesota would be willing to take part in a comparability study and would be open to having the test mode for each student be determined randomly.

The testing window for the study was April 27–May 15, 2009. The dates were chosen to take place after the MCA-II census administration window closed but before the release of test scores. Because students participating in the study had already taken the census test with embedded GRAD, taking part in the comparability study represented an extra chance to pass the GRAD standard. The additional opportunity to pass the GRAD provided motivation for school and student participation. Because the study took place before MCA-II scores were reported, the students in the study did not know if they had passed the GRAD standard by virtue of their scores on the MCA-II or the embedded GRAD.

Online and paper versions of a test form were created that matched all the psychometric and content constraints contained in the GRAD test specifications. The test form contained 40 multiple-choice test questions. Although the online and paper versions used the same test questions, the items were formatted differently in the two testing modes. The particular format used for each mode mimicked the formatting used on either the MCA-II (paper) or GRAD retests (online). For example, the paper version used Frutiger font, whereas the online test used Verdana font. Also, formatting differed on how the questions appeared on screen versus on paper. On paper, four questions would often appear on a single page, with two questions side by side. This format, used in the paper MCA-II, is easy to read and makes efficient use of the page. However, splitting the page vertically into two sections narrows the space available for each line of text and results in questions that are narrower and taller.

For the online version, it was desired to minimize or eliminate scrolling, in order that the entire question and any associated graphics could be seen at one time. Thus, on the online

version questions appeared one at time and tended to be wider and shorter. The purpose of the comparability study was to investigate whether the differences associated with administration mode, including formatting differences, would impact student performance on the test.

Sampling Plan

The strength of the randomized groups design is derived from obtaining representative and randomly assigned samples. Because proper sampling is critical in this design, the psychometric groups from MDE and Pearson worked in conjunction to formulate a sampling plan. The goal of the sampling plan was to identify a set of schools and districts within the state whose students would form a sample broadly representative of the state's eleventh-graders in terms of gender, ethnicity, school size, economic status, and geographic location.

A stratified sampling approach was employed using school geographic location (urban, suburban, rural) and free/reduced-price lunch program eligibility percentage (hereafter denoted FRP). School location was chosen as a factor in recognition that computer familiarity could vary between urban, suburban, and rural schools. The FRP variable ("Yes" if the student participated in free/reduced-price lunch, "No" otherwise) served as an indicator of economic status. Economic status is believed to be associated with the amount of resources a school has to buy computer equipment as well as the likelihood that students will have acquired computer familiarity at home. Thus, higher percentages of FRP students might be an indication of less computer familiarity.

The school location (urban, suburban, rural) and FRP (<20% FRP, 20–40% FRP, >40% FRP) factors were crossed, resulting in nine cells, or strata. The schools from the state were grouped into the nine strata. In some of the larger strata, schools were further subdivided based on their percentage of minority enrollment. Within each stratum/substratum, MCA-II Grade 11 Mathematics scale score means and standard deviations for 2008 were calculated, and a target number of students was calculated proportional to the total number of the students in the stratum. (The proportion was calculated to result in a total sample of 2,500.)

Cost and logistic considerations limited the number of students and the number of schools that could be sampled. As a consequence, random sampling within strata was not feasible. Instead, schools were sampled purposively within each stratum so as to cover a range of school sizes (when the target sample size permitted), reflect ethnic diversity within the stratum, and yield MCA-II means and standard deviations close to those of the stratum.

Based on the original sample selection, 36 schools were contacted and invited to participate in the study. Of these, 23 agreed to participate (64%). Upon receipt of a refusal, an invitation was extended to a similar school from the same stratum/substratum (when possible) or an adjacent stratum. At the end of the recruitment process, 34 schools (of 49 invited, or 69%) had agreed to participate; two of these schools later dropped out.

A comparison of the state target and school sample characteristics for 2008 suggested that the method resulted in a sample that would closely track state characteristics, except for oversampling of FRP-eligible students. The oversampling of FRP students was intentional, as these are the students for whom computer familiarity is most questionable. The increased sample size of the FRP group increases the chances of detecting a mode effect for these students, should one exist. MDE considered the advantage of increased sensitivity to outweigh the disadvantages associated with deviation from representative sampling.

Mode Assignment

The student roster for each of the participating schools was used to randomly assign students to testing conditions: either paper test or computer test. Schools were notified which condition was assigned for each student. After administration of the test was completed and data were available, the assessment of students under their assigned condition was verified. MDE reviewed the data for any misadministration issues. These issues included testing under the wrong mode, participation by students not on the roster, students choosing not to participate, etc. Extra paper copies of the form were given to the schools so that students who were not on the roster could take the test and have an

opportunity to pass the GRAD standard. However, data from students not on the roster or who were misassigned were excluded from the study. Also excluded were students who did not have valid scores (e.g., they did not attempt the test).

The rules used to define the valid test scores were the same rules that will be applied to the online operational GRAD retests. Table 1 gives the number of students who took the online or paper version of the form and the number of students with valid scores taking part in the study. The main reason that the paper version had a greater number of testtakers is that paper forms were given to students who were not on the precoded roster. The numbers in the Total Participants with Valid Scores category only consider those students who were on the roster and who took the mode that they were assigned to.

Table	1.	Study	Sample	Size
-------	----	-------	--------	------

Total Test Takers		Total Participants with Valid Scores		
Online	Paper	Online	Paper	
1,139	1,184	1,036	1,035	

Sample Characteristics

As described above, the overall sample, as well as the online and paper subsamples, was intended to be representative of statewide eleventh-grade students, with FRP deliberately oversampled. Table 2 presents the MCA-II test scores of the obtained sample and the test scores from the 2008 statewide population. Additionally, the column labeled 2008 Scores of Sample Schools shows the expected test scores of the intended sample using 2008 test scores from the schools in the sample.

	2008 Statewide Scores	2008 Scores of Sample Schools	2009 MCA-II Scores of Obtained Sample			2009 Statewide Scores
			Total	Online	Paper	
Sample Size	2500 (target)	2499	2026	1011	1015	62379
Mean MCA-II Score	1141.09	1140.01	1144.77*	1144.77*	1144.77*	1144.54
Standard Deviation of MCA-II Score	20.59	19.82	18.35	18.49	18.21	20.19
Proportion Meets Standard or Above	0.34	0.31	0.42	0.42	0.42	0.42

Table 2. Sample Versus Target Test Scores

*In a somewhat remarkable coincidence, the mean MCA-II scores of the Online and Paper samples matched exactly to two decimal places.

One point to note from Table 2 is the difference between test scores in 2008 and 2009 for the sampled schools. Although this difference might imply that the obtained sample is of higher proficiency than what was desired or expected, another factor that must be taken into account is that 2009 was the first year that the MCA-II test (along with the embedded GRAD test) counted for graduation requirements. A motivation effect could explain the higher scores of the obtained samples test scores. The last column in Table 2 gives the statewide 2009 test scores. As can be seen when comparing the 2009 average (1,144.54) versus the 2008 average (1,141.09), the 2009 population scored higher, perhaps due to the aforementioned motivation effect. The results in Table 2 suggest that the obtained sample was comparable to the state population.

Student Catagory	2008	2008		2009	
Student Category	Statewide	Sample Schools	Ol	otained Sam	ple
			Total	Online	Paper
Free/Reduced- Price Lunch Eligible (FRP)	23%	29%	31%	30%	33%
Limited English Proficiency	2%	1%	4%	4%	4%
Special Education	10%	10%	9%	10%	8%
Black, Non- Hispanic	7%	9%	10%	8%	11%
American Indian	2%	4%	2%	3%	2%
Asian/Pacific Islander	5%	6%	9%	9%	9%
Hispanic	3%	3%	4%	4%	4%
White, Non- Hispanic	83%	78%	75%	75%	75%
All Non-White	17%	22%	25%	25%	25%

Table 3. Sample Versus Target Demographic Characteristics

Table 3 presents the demographic characteristics of the obtained sample and the targets. The 2008 Statewide column gives the demographic characteristics of the state for that year. These proportions served as the targets for the sampling plan. Comparing the 2008 Statewide and 2008 Sample Schools columns shows that the selected schools were generally representative of the state. The most notable exception was the FRP category, which was initially oversampled. The columns under the heading 2009 Obtained Sample give the demographic characteristics of the final sample.

The table shows that, overall, the obtained sample matched the targets well. Ethnic group departures from targets occurred primarily with African Americans and Asians being slightly overrepresented and Whites being underrepresented in the sample. It is likely that

the observed minority differences from the targets were due to the intentional oversampling of the FRP group.

One threat to the internal validity of the study is the potential for differential participation rates in the two testing modes that may be confounded with achievement differences. Even though students were assigned at random to the two testing conditions, not all students participated, and participation might be associated differently with student achievement in the two modes. In order to evaluate this threat, statistical comparisons were conducted between participants who achieved valid scores in the comparability study. Comparisons of demographic characteristics (gender, ethnicity, ELL status, economic status, geographic region, and special education status) of the students in the two test modes indicated no significant differences between the two groups.

Table 4 gives summary statistics of the GRAD census raw scores for those students in the comparability study who took the 2009 April census GRAD. The summary statistics are broken out by various demographic groups for the online and paper samples. Under the assumption of random assignment to groups, no differences on the GRAD Census score mean would be expected. Results from Table 4 suggest that the groups taking the two modes were quite comparable. Only small differences were found between the paper and online groups, and a series of t-tests showed no statistically significant difference at p < .05. For a typical hypothesis testing setting, conducting a large number of t-test comparisons is not advisable, due to the likelihood of finding spurious effects. In the present situation, however, the high degree of power obtained through the t-test comparisons provides strong support for the contention that the minor differences observed between online and paper groups can be attributed to sampling error. These results suggest that factors leading to participation did not subvert the randomization process.

	Assigned			Standard	Proportion
Group	Condition	Ν	Mean	Deviation	Passing
A 11	Online	1011	27.92	7.22	.58
All	Paper	1016	28.01	7.14	.58
Famala	Online	490	27.46	7.17	.55
Female	Paper	519	27.52	7.16	.55
Mala	Online	521	28.35	7.24	.61
Iviale	Paper	497	28.52	7.09	.60
Dissle New Hirmania	Online	81	22.57	7.27	.27
Black, Non-Hispanic	Paper	105	23.13	7.88	.37
American Indian	Online	31	23.65	7.82	.35
American Indian	Paper	20	23.95	6.39	.35
Asian/Pacific	Online	88	27.80	6.78	.56
Islander	Paper	87	28.40	6.14	.54
Hispanic	Online	35	23.31	8.72	.40
	Paper	35	25.66	7.50	.40
White Non Hignoria	Online	776	28.87	6.78	.63
White, Non-Hispanic	Paper	769	28.85	6.82	.62
All Non White	Online	235	24.78	7.72	.41
All INOII- willte	Paper	247	25.41	7.48	.43
Free/Reduced-Price	Online	287	25.03	7.71	.43
Lunch Eligible (FRP)	Paper	332	25.58	7.62	.45
	Online	27	21.59	7.67	.22
ELL	Paper	23	23.87	7.29	.26
	Online	100	20.49	7.94	.22
Special Education	Paper	80	19.46	7.91	.18
	Online	435	28.15	7.15	.58
Kurai	Paper	447	28.23	6.97	.58
Suburban	Online	413	28.58	7.00	.64
Suburban	Paper	413	28.52	7.01	.61
T-l	Online	163	25.61	7.51	.44
Urban	Paper	156	26.06	7.66	.47

Table 4. GRAD Census Raw Score Mean and Standard Deviations and Pass Rates

Two other threats to the validity of the study were also addressed. The first of these concerned the impact of student motivation on test score results. It was anticipated that some students would not be fully motivated participants in the comparability study testing. To reduce the impact of motivation-related measurement error, several steps were taken. First, scores of students who scored below chance on the comparability study administration of the GRAD were excluded from the subsequent analyses of mode effects. This rule led to the exclusion of 26 students (1.3% of the participants). Two

additional students who had scored below chance on the census GRAD administration were excluded from analyses that used the census GRAD score as a covariate.

Next, the consistency of individuals' GRAD raw scores on the census administration and the comparability study administration was examined. The rationale for this was that extremely large score differences (i.e., 15 raw score points or more in either direction) were likely attributable to motivation differences; inclusion of those scores in the analysis could hamper our ability to detect more subtle mode effects. To the extent that substantial score changes were equally frequent and of equivalent size among participants in the two modes, removal of those cases should not importantly bias the results. During the course of this examination, our attention was drawn to a school that appeared to have disproportionate representation among students with large score changes; in particular, score decreases among online mode students were observed. This suggested the possibility of another threat to the validity of the study: systematic site differences in the implementation of the study conditions.

In order to evaluate the extent of such heterogeneity in mode effects on the GRAD score across sites, we employed a mixed effects model approach that included random intercept and mode effect terms, as well as fixed effects (GRAD census score, gender, ELL status, economic status, special education status, ethnicity, and geographic classification). The estimated variance of the random mode effect was 1.81 (SE = .85), and significantly different from zero (p < .01). This means that even after adjustment for demographic variables and census GRAD score, variation in mode effects across sites was greater than that attributable to sampling variability.

Inspection of empirical Bayes estimates of site-specific mode effects indicated that the site in question was an extreme outlier (see Figure 2). In order to further evaluate the reasonableness of the suspect site's data, a bootstrapping procedure was used. Samples of online and paper mode respondents (with *n* counts equal to those of the site) were drawn at random from the study population, and mean differences between online and paper mode samples were calculated. In only one of 1000 samplings did the mean difference exceed that observed at the target site.





Given the statistical evidence for aberrance of this site, MDE sought further information about the implementation of the study there. Review of records indicated that the school had lagged in offering the online GRAD (it occurred a full week after the paper administration) and that the online administration had been hurriedly scheduled after repeated calls from MDE project staff urging the school to complete the study. In view of the problematic circumstances surrounding the school's implementation of the study, and the statistical evidence for anomaly, the data from all students (n = 99) at that site were excluded from the subsequent comparability study analyses.

Of the remaining participants with both census and comparability study GRAD scores above chance, 20 (1.05%) had raw score changes of 15 points or more; 17 of these were score declines from census administration to the comparability study administration. Representation from the online and paper mode students was relatively balanced: 11 online students and 9 paper students overall. Of the 3 students manifesting raw score gains of 15 points or greater, 2 were from the online mode. In view of this balance, it seemed that the exclusion of these 20 participants would be unlikely to bias the study results but could serve to improve sensitivity of the analyses. Therefore, they were excluded from subsequent analyses of mode effects.

Comparability Results

Testing mode comparability results are examined in terms of mean differences, score distributions, and statistical modeling techniques. Mean differences are given in Table 5, which displays raw score means, raw score standard deviations, and the passing rate for the overall sample and various subgroups. The passing score was defined as a raw score of 28 for both the online and paper versions. This corresponds to the pre-equated cut score for the test form (i.e., the cut score that would be used if the form were given as a GRAD retest). Note that sample sizes may vary from those given in Table 4 due to the removal of outliers discussed above and the fact that not every student in the study had a GRAD Census score.

Table 5 shows that the differences between modes varied quite a bit across subgroups, but in most cases means and passing rates from the paper group were higher. Exceptions to this pattern were observed in the American Indian and Special Education subgroups, both of which had small sample sizes. Further mode comparisons are given in Table 6, which presents the frequency distributions of the online and paper samples. It should be noted that the Table 6 counts include cases which scored below chance on the comparability study GRAD; these cases are excluded in the other mode effect analyses. The table is consistent with the data in Table 5, which showed that the paper test had higher scores. Table 6 shows that the difference between modes was manifested across most of the score distribution.

				Standard	Proportion
Group	Mode	Ν	Mean	Deviation	Passing
A 11	Online	959	27.03	7.72	.53
AII	Paper	968	27.37	7.88	.56
Famala	Online	466	26.73	7.38	.51
remaie	Paper	495	26.87	7.81	.53
Mala	Online	493	27.32	7.71	.54
Male	Paper	473	27.89	7.92	.59
Dlask Non Himonia	Online	82	22.27	7.06	.22
Black, Non-Hispanic	Paper	107	22.67	7.60	.29
A morizon Indian	Online	31	22.65	8.50	.29
American Indian	Paper	20	21.20	8.70	.25
Asian/Pacific	Online	92	26.27	7.48	.47
Islander	Paper	92	27.99	6.98	.59
Hignania	Online	41	22.49	8.17	.32
пізрапіс	Paper	38	24.97	7.90	.39
White Non Hignoria	Online	713	28.13	7.40	.59
white, won-mispanic	Paper	711	28.30	7.66	.61
All Non White	Online	246	23.85	7.78	.34
All Ivon- winte	Paper	257	24.80	7.90	.41
Free/Reduced-Price	Online	289	24.05	8.00	.37
Lunch Eligible (FRP)	Paper	330	25.14	7.93	.42
	Online	39	19.28	6.27	.15
ELL	Paper	39	23.08	8.32	.31
	Online	87	19.94	8.37	.22
Special Education	Paper	77	19.43	7.29	.18
	Online	428	27.43	7.65	.55
Rural	Paper	438	27.59	7.70	.56
	Online	364	27.65	7.59	.57
Suburban	Paper	370	27.80	8.07	.58
TT P	Online	167	24.68	7.81	.37
Urban	Paper	160	25.76	7.74	.48

Table 5. Raw Score Mean and Standard Deviations and Pass Rates

	Online			Paper		
Raw			Cumulative			Cumulative
Score	Frequency	Percent	Percent	Frequency	Percent	Percent
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	1	0.10	0.10
5	1	0.10	0.10	1	0.10	0.20
6	3	0.31	0.41	1	0.10	0.31
7	5	0.51	0.93	3	0.31	0.61
8	3	0.31	1.24	4	0.41	1.02
9	9	0.93	2.16	5	0.51	1.53
10	4	0.41	2.57	6	0.61	2.15
11	5	0.51	3.09	16	1.64	3.78
12	16	1.65	4.74	16	1.64	5.42
13	19	1.96	6.69	13	1.33	6.75
14	16	1.65	8.34	19	1.94	8.69
15	25	2.57	10.92	19	1.94	10.63
16	19	1.96	12.87	32	3.27	13.91
17	24	2.47	15.35	17	1.74	15.64
18	30	3.09	18.43	24	2.45	18.10
19	28	2.88	21.32	22	2.25	20.35
20	23	2.37	23.69	30	3.07	23.42
21	32	3.30	26.98	30	3.07	26.48
22	27	2.78	29.76	25	2.56	29.04
23	34	3.50	33.26	20	2.04	31.08
24	42	4.33	37.59	35	3.58	34.66
25	37	3.81	41.40	33	3.37	38.04
26	29	2.99	44.39	33	3.37	41.41
27	36	3.71	48.09	34	3.48	44.89
28	32	3.30	51.39	46	4.70	49.59
29	52	5.36	56.75	43	4.40	53.99
30	41	4.22	60.97	49	5.01	59.00
31	49	5.05	66.01	43	4.40	63.39
32	53	5.46	71.47	46	4.70	68.10
33	46	4.74	76.21	50	5.11	73.21
34	39	4.02	80.23	46	4.70	77.91
35	48	4.94	85.17	49	5.01	82.92
36	44	4.53	89.70	58	5.93	88.85
37	38	3.91	93.61	44	4.50	93.35
38	31	3.19	96.81	33	3.37	96.73
39	20	2.06	98.87	21	2.15	98.88
40	11	1.13	100.00	11	1.12	100.00

Table 6. Frequency Distribution of Scores by Mode

Although differences were observed between modes, this does not necessarily imply that the sample differences are statistically reliable, or that the paper form was easier than the online form. Critical in the analysis of results from any experimental study is an acknowledgment that differences in means between two groups can be attributed to either systematic effects due to the experimental conditions or to random error in estimating the means due to sampling. Even a representative sample can only estimate the true population mean. If the current study were repeated with a new sample of schools representative of the statewide student population (with FRP oversampling), group means would differ from those found here simply due to sampling variation. The difference between the sample mean and the population mean is called *sampling error*. To help sort out differences that are statistically reliable from those that are due to sampling error, statistical tests of significance are employed.

Statistical tests of significance could have been performed on the group differences for each of the groups given in Table 5. However, conducting a large number of independent statistical tests inflates the probability of making a Type I error (i.e., incorrectly concluding that a testing mode difference exists). A second, related issue is that simple mean comparisons (e.g., t-tests) may be misleading when the units of observation (students) are clustered in schools. Appropriate statistical procedures need to reflect that clustering when evaluating testing mode effects. Furthermore, when appropriate covariates related to the dependent variable are included in the analyses, they can allow for more sensitive statistical tests by reducing error variance.

In order to address these concerns, a hierarchical modeling approach was taken by comparing nested, linear mixed models that reflected student clustering in schools and included demographic and prior GRAD score covariates. The basic approach is to sequentially compare the ability of pairs of models to predict student performance on the GRAD comparability test administration. The two models differ only in that the more complex model includes additional terms reflecting mode differences. If the more complex model provides a statistically significant better prediction than the simpler model, then the inference is that there is evidence to support the existence of additional test mode effects, and these would be probed in greater detail. Conversely, if the more

complex model fails to demonstrate better prediction (fit) than the simpler model, then the inference is that the evidence does not support test mode effects.

The composition of the three models tested is summarized in Table 7. It can be seen that school is considered a random effect in both models. Census GRAD score is included as a covariate in all the models, as are demographic covariates: gender, geographic area (urban, suburban, rural), English language learner (ELL) status, special education (SpEd) status, economic (free or reduced-price lunch; FRP) status, and ethnicity. The mode effects included in the most complex model, Model 1, reflect overall mode impact, as well as mode impact specific to the subgroups defined by the demographic variables. The mode x GRAD-census-score interaction term reflects mode effects that vary, depending on achievement level. Model 2 drops all mode-related interaction terms, while retaining the overall mode effect term. Model 3 drops all mode effects.

The three models were fit using SAS PROC MIXED, with full maximum-likelihood estimation. Comparison of model fit indices (deviance) for nested models provides a basis for statistical comparison of the models, as the differences are asymptotically distributed as χ^2 with degrees of freedom equal to the difference in degrees of freedom for the two models. Inspection of the results for models 1 and 2 in Table 7 indicates that there is not a statistically significant effect associated with testing mode interactions ($X^2 = 7.0$, df = 11, p > .50). That is, significant mode x demographic or mode x pretest score interaction effects on GRAD score were not detected. Similarly, comparison of Model 2 versus Model 3 provides no evidence that a test mode effect exists ($X^2 = 0.9$, df = 1, p > .30).

	Random	Fixed Effects:	Fixed Effects:	Model	Model
Model	Effect	Covariates	Mode Related	Fit	DF
1. Test Mode Interaction with Covariates	School	GRAD Census score Gender Geographic Area ELL Status SpEd Status FRP Status Ethnicity	Test Mode Mode x GRAD Census Mode x Gender Mode x Geographic Mode x ELL Status Mode x SpEd Status Mode x FRP Status Mode x Ethnicity	10656.5	26
2. Test Mode and Covariates	School	GRAD Census score Gender Geographic Area ELL Status SpEd Status FRP Status Ethnicity	Test Mode	10663.5	15
3. Covariates Only	School	GRAD Census score Gender Geographic Area ELL Status SpEd Status FRP Status Ethnicity		10664.4	14

 Table 7. Comparison of Nested, Linear Mixed Models of GRAD Comparability

 Study Scores

A second set of analyses tested for mode effects on pass/fail outcomes. As before, the fit of pairs of nested logistic models was compared, this time using SAS PROC NLMIXED. The results of these analyses are reported in Table 8. As with the test score analyses, the effects associated with test mode interaction terms did not reach statistical significance (Model 1 versus Model 2 $X^2 = 11.1$, df = 11, p > .30). Similarly, comparison of fit statistics for Model 2 versus Model 3 indicated no significant effect for test mode ($X^2 = 2.4$, df = 1, p > .10).

	Random	Fixed Effects:	Fixed Effects:	Model	Model
Model	Effect	Covariates	Mode Related	Fit	DF
1. Test Mode Interaction with Covariates	School	GRAD Census score Gender Geographic Area ELL Status SpEd Status FRP Status Ethnicity	Test Mode Mode x GRAD Census Mode x Gender Mode x Geographic Mode x ELL Status Mode x SpEd Status Mode x FRP Status Mode x Ethnicity	1237.4	25
2. Test Mode and Covariates	School	GRAD Census score Gender Geographic Area ELL Status SpEd Status FRP Status Ethnicity	Test Mode	1248.5	14
3. Covariates Only	School	GRAD Census score Gender Geographic Area ELL Status SpEd Status FRP Status Ethnicity		1250.9	13

 Table 8. Comparison of Nested, Mixed Logistic Models of GRAD Comparability

 Study Pass/Fail Outcomes

Although use of the GRAD census score as a covariate in these analyses increased their statistical power by reducing error variance (the observed product-moment correlation between the two GRAD raw scores was .84), doing so excluded data from 41 students who had not taken the MCA-II/GRAD census test. Of the excluded cases, 31 (75.6%) were ELL students. This result is not surprising, given that a substantial portion of ELL students take the Mathematics Test for English Language Learners (MTELL) instead of the MCA-II. The MTELL—a reduced-language, accommodated version of the MCA-II—does not include an embedded GRAD test. In order to address our concerns that exclusion of cases with no census GRAD score might have biased the results, additional analyses, similar to those reported in Tables 7 and 8 but without the census GRAD score covariate, were conducted. (Other modifications involved addition of some two-way demographic interaction terms and the inclusion of students who had large score changes between the census test and the comparability study.) The results of those analyses, reported in Tables 9 and 10, led to the same conclusions found in the previous analyses.

For the linear mixed models of GRAD raw score reported in Table 9, a comparison of models 1 and 2 indicated no statistically significant effect for mode of administration in interaction with demographic factors ($X^2 = 9.4$, df = 10, p > .50). A comparison of models 2 and 3 indicated no simple mode effect ($X^2 = 1.1$, df = 1, p > .25). Similarly for the dichotomous pass/fail outcome, as reported in Table 10, comparisons of models 1 and 2 ($X^2 = 9.6$, df = 10, p > .50) and models 2 and 3 ($X^2 = 2.6$, df = 1, p > .10) effects associated with testing mode were not identified as statistically significant.

	Random	Fixed Effects:	Fixed Effects:	Model	Model
Model	Effect	Covariates	Mode Related	Fit	DF
1. Test Mode Interaction with Covariates	School	Gender Geographic Area ELL Status SpEd Status FRP Status Ethnicity Ethnic X Geographic Gender X FRP	Test Mode Mode x Gender Mode x Geographic Mode x ELL Status Mode x SpEd Status Mode x FRP Status Mode x Ethnicity	13011.4	33
2. Test Mode and Covariates	School	Gender Geographic Area ELL Status SpEd Status FRP Status Ethnicity Ethnic X Geographic Gender X FRP	Test Mode	13020.8	23
3. Covariates Only	School	Gender Geographic Area ELL Status SpEd Status FRP Status Ethnicity Ethnic X Geographic Gender X FRP		13021.9	22

 Table 9. Comparison of Nested, Linear Mixed Models of GRAD Comparability

 Study Scores (No GRAD Census Covariate)

	Random	Fixed Effects:	Fixed Effects:	Model	Model
Model	Effect	Covariates	Mode Related	Fit	DF
1. Test Mode Interaction with Covariates	School	Gender Geographic Area ELL Status SpEd Status FRP Status Ethnicity Black X Urban Gender X FRP	Test Mode Mode x Gender Mode x Geographic Mode x ELL Status Mode x SpEd Status Mode x FRP Status Mode x Ethnicity	2335.9	25
2. Test Mode and Covariates	School	Gender Geographic Area ELL Status SpEd Status FRP Status Ethnicity Black X Urban Gender X FRP	Test Mode	2345.5	15
3. Covariates Only	School	Gender Geographic Area ELL Status SpEd Status FRP Status Ethnicity Black X Urban Gender X FRP		2348.1	14

 Table 10. Comparison of Nested, Mixed Logistic Models of GRAD Comparability

 Study Pass/Fail Outcomes (No GRAD Census Covariate)

As part of these analyses, some additional, finer-grain probing of group-specific mode effects was done, but the failure to reject the omnibus hypothesis of no mode effects cautions against placing much reliance on any inferences that might be made. There was some suggestion that ELL students may have performed less well on the online test, but the impact of such a group-specific mode effect, if real, would be attenuated by the provision of state law that exempts ELL students with less than four years attendance in Minnesota schools from having to pass the GRAD. If a test administrator believes that a student lacks adequate computer skills to complete the online GRAD test, the test administrator may wish to provide additional computer training or consider use of the alternate paper form.

The mode comparisons in Table 5 through Table 10 focused on raw score differences. Because the GRAD retests are pre-equated using the 3PL model, it is important to also examine whether the IRT parameter estimates show any influence from the administration mode. In this study, the overall mode effect on the IRT parameter estimates was investigated by looking at the raw score to theta score transformation at the cut score. To obtain an estimate of equating error for the sample size at hand, a bootstrap procedure was used. The bootstrap is a statistical process whereby repeated sampling is done from the observed data (with replacement) so that empirical standard errors can be obtained. The bootstrapping approach given here is based on the one described in Way, Davis, and Fitzpatrick (2006). In their study, they recommended using the following rule: If the statistic of interest is within two empirical standard errors of the null value, then the observed difference is judged to be not statistically significant.

Five hundred bootstrap samples were created for each of the online and paper versions of the test. To create a single online bootstrap sample, a random sample (with replacement) was drawn from the actual online sample. The size of each bootstrap sample was chosen to equal the sample size of the real sample (e.g., 1036 for online) so that the calculated bootstrap standard errors would be based on the appropriate sample size. The process was repeated for the other online bootstrap samples. The paper bootstrap samples were created similarly.

The bootstrap samples were used in an analysis where the online IRT parameter estimates were scaled to the paper estimates. Since the samples taking the online and paper versions were randomly equivalent, if no mode effect is present, the resulting scaling should differ from the identity transformation only by sampling error. If a mode effect is present, however, the resulting scaling would be expected to differ from the identity function. For each bootstrap sample drawn, the following steps were conducted:

- Step 1: MULTILOG was used to separately calibrate the online bootstrap data and the paper bootstrap data.
- Step 2: The Stocking-Lord equating procedure was used to find the scaling constants that best placed the item parameter estimates from each online bootstrap sample on the scale of the corresponding paper bootstrap sample.

Table 11 shows the results. The slope and intercept values in the table are the Stocking-Lord scaling constants obtained from scaling the real data online IRT parameters to the real data paper parameters. The bootstrap standard error values given are the standard deviation of the slope and intercept parameters across the five hundred bootstrap samples. The bootstrap standard deviation may be thought of as an estimate of the standard error of the parameter. The slope and the intercept values are approximately one standard error from the identity function values and are therefore judged to be not statistically significant following the rule described above.

Table 11. Scaling Transformation and Bootstrap Standard Error

Slope	Bootstrap Standard Error	Intercept	Bootstrap Standard Error
1.05	.05	.05	.05

Conclusion

The GRAD Mathematics comparability study benefited from implementing the randomized groups design. This design is powerful and avoids confounds that can occur in other designs. The study's sample was found to be generally representative of the statewide Grade 11 students, both in terms of demographics and test scores. An important exception was an intentional oversampling of FRP students.

Because comparability work that Pearson has done in a number of other states has not found consistent results, it was difficult to make a prediction for the outcome of the current study. A variety of analyses were utilized to investigate whether the GRAD Math assessment is susceptible to mode effects. Small mean differences for various subgroups generally favored the paper test. However, a hierarchical modeling approach found no statistically reliable differences between modes. Also, results from a bootstrap IRT analysis scaling the online test to the paper test found scaling constants that did not statistically differ from the identity scaling function. Taken as a whole, the statistical analysis failed to find reliable evidence of a mode effect. The GRAD mathematics form selected for the study was not unusual in any way and was built to the same specifications as the other GRAD forms. Because the items in the pool are fairly consistent in format, it seems unlikely that the results of the study would have changed importantly had different items been employed. The recommendation from the study was that no statistical adjustment be made for scaling the GRAD Mathematics online tests. The MDE determined that no such adjustment would be made.

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.
- Way, W.D., Davis, L.L., and 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 for Measurement in Education, San Francisco, CA.