Maintaining Score Equivalence as Tests Transition Online: Issues, Approaches and Trends

Walter D. Way
Chow-Hong Lin
Jadie Kong

Pearson

Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY, March 2008

Acknowledgements

The authors are grateful to the following state testing personnel for permitting us to share the results of studies done for their programs:

Gloria Zyskowski – Texas Education Agency

Joseph Martineau – Michigan Office of Educational Assessment and Accountability

Elizabeth Jones – South Carolina Department of Education

Roberta Alley – Arizona Department of Education

Marty Kehe – Maryland Department of Education

Maintaining Score Equivalence as Tests Transition Online: Issues, Approaches and Trends

Introduction

Computer-based testing has been around for more the 25 years and its rich history already includes both successful and not-so-successful applications (McDonald, 2007; Mayfield, 2002). In recent years, the expansion of large-scale state assessments in the United States in response to the *No Child Left Behind Act of 2001* (NCLB) has resulted in a new chapter in the history of online testing. Many state education departments are exploring or have implemented online assessments as part of their statewide assessment programs. Within the context of how technology is evolving in education, the idea of online testing is compelling (Bennett, 2002). However, as most states pursue online testing, they find that not all schools have the infrastructure and equipment to test all of their students online. For this reason, paper and online versions of the same tests are typically offered side-by-side. Any time both paper-based and online assessments coexist, professional testing standards indicate the need to address comparability of results across paper and online mediums (APA, 1986; AERA, APA, NCME, 1999, Standard 4.10).

The comparability of online versus paper test scores has been studied for about as long as computer-based testing has been in existence. Literature reviews were reported by Mazzeo and Harvey (1988), who found mixed results, and Mead and Drasgow (1993), who concluded that there were essentially no mode differences in examinee scores for power tests. A number of studies have indicated minimal mode differences in a variety of settings (Kim & Huynh, in press; Kim & Huynh, 2007; Poggio, Glassnapp, Yang, & Poggio, 2005; Hetter, Segall & Bloxom, 1997; Bergstrom, 1992; Spray, Ackerman, Reckase, & Carlson, 1989). However, some recent studies have also reported lower performance for students testing online. Such findings have been reported for reading tests with items requiring text scrolling (Way, Davis & Fitzpatrick, 2006; O'Malley et al., 2005; Pommerich, 2004; Bridgeman, Lennon, & Jackenthal, 2003; Choi & Tinkler, 2002), for timed achievement tests (Ito & Sykes, 2004), and for math items that require graphing and geometric manipulations (Keng, McClarty & Davis, 2006). On the other hand, recent meta-analyses by Wang et al. (2008; 2007) of K-12 studies concluded that administration mode had no statistically significant effect on either student reading or mathematics achievement scores.

These mixed findings promote ambiguity in considering the need to address mode comparability as tests transition online. For a policy-maker interested in introducing online assessments for a high-stakes K-12 testing program, it is difficult to sort out these various studies. Moreover, while empirical studies addressing comparability seem like a good idea, carrying out controlled experiments within the context of a statewide assessment program is difficult if not impossible. For this reason, quasi-experimental designs are often the only viable research option.

In our work with state departments of education, we have depended almost exclusively upon quasi-experimental designs to evaluate the comparability of paper and online assessments. In particular, we have adopted a comparability design we refer to as Matched Samples Comparability Analyses (MSCA; Way, Um, Lin & McClarty, 2007; Way, Davis & Fitzpatrick, 2006). This approach utilizes covariates to create comparable online and paper groups, summarizes comparability results in the context of test equating, and can result an alternate online group score conversion table to correct for mode effects. The design requires no special assignment to conditions within schools, and can be executed as an extension of the general test equating process.

The purpose of this paper is to summarize studies that we have conducted with K-12 state departments of education using the MSCA method. This includes 46 studies across five different states. In addition, we discuss a number of trends that we see emerging from the K-12 comparability studies done to date and discuss the issues that face efforts to monitor the comparability of online and paper versions of tests over time.

Description of MSCA

MSCA involve a bootstrap method designed to establish raw to scale score conversions by equating the online form to the paper form. The method also estimates bootstrap standard errors of the equating to assist in interpreting differences between the online and paper score conversions (c.f., Kolen & Brennan, 2004, p. 232-235). In MSCA applications, the samples of online and paper test-takers are typically self-selected. Usually the paper sample is much larger than the online sample, but this is not a requirement for using the method. The method assumes that both samples include previously or concurrently obtained test scores in a content area correlated with the content area under study, as well as additional demographic information (e.g.,

gender, ethnicity). The method proceeds by drawing a sample of students testing online and matching them to students taking the paper form (or vice versa) so that the two samples have identical profiles of previous test scores and demographic variables. This matching can be based strictly on a score variable or on a table of score levels by demographic categories (e.g., 20 equally-sized score categories by 2 gender groups by 4 ethnic groups). The two samples are then equated under the assumption of randomly equivalent groups. The procedure is repeated for some number of replications (e.g., 100 or 500) and the equating results are summarized over replications. The means of the equated online score conversions at each raw score point are compared to the paper score conversions, and the standard deviations over replications (i.e., the bootstrapped standard errors) are used to interpret the comparisons.

Within a bootstrap iteration, the matched samples can be equated using any method, including IRT and non-IRT approaches. We have most often applied the Rasch or Rasch partial credit models as part of MSCA, but we have also used 3PL and GPC models for programs where these models are employed in practice. Way, Um, Lin, and McClarty (2007) compared the use of Rasch, 3PL, and equipercentile methods as part of MCSA with simulated data and found little difference between these approaches.

In recent applications of the method we have utilized linear regression to obtain predicted score levels so that we can incorporate multiple predictor scores into the matching process. For example, in some studies we have utilized previous scores on ELA, Math, Science, and Social Studies tests to predict scores on online and paper versions of an ELA test using the following regression model:

$$\hat{Y}_{predicted\_score} = \beta_0 + \beta_1 X_{1(\mathrm{Pr}ev\_ELA)} + \beta_2 X_{2(\mathrm{Pr}ev\_Math)} + \beta_3 X_{3(\mathrm{Pr}ev\_Science)} + \beta_4 X_{4(\mathrm{Pr}ev\_SocialSt.)}$$

We then divide the predicted ELA score into 20 equally-sized score categories and use these categories to obtain strata for the bootstrap sampling. For additional information about the MSCA approach, see Way, Davis & Fitzpatrick (2006) and Way, Um, Lin, and McClarty (2007).

The bootstrap approach makes it possible for the user compare the paper and online groups within iteration and also to aggregate the results over iterations. Groups can be compared in terms of mean differences between scores by calculating a standardized mean difference statistic across iterations 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 replications and SE_{diff} and is the bootstrap standard deviation of the differences over the replications.

The effect size between two group mean ability estimates at each replication (see Cohen, 1992) can be calculated within iteration using 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 can then be calculated based on the averages of the effect sizes over replications.

Finally, proportion correct or *p-values* for each operational item in the online and paper versions of a test and differences between the p-values can also be calculated during each bootstrap iteration. The p-value differences can then be averaged over the replications for each item and a standardized mean difference statistic for each item can be calculated using the *Zdif* statistic described above. Based on advice from state technical advisors, in some applications we have used an *adjusted critical z-value* that multiplies the conventional critical value of z = 1.96 (at an alpha level of 0.05) by a factor of the bootstrap sample size over a base sample size of 500 (H. Huynh., personal communication, August 17, 2006).

MSCA has a number of similarities to a method referred to as propensity scoring (Rosenbaum & Rubin, 1985). This method matches students in two groups on a single variable that represents a composite of a group of covariates. The propensity score is a weighted sum of predictor variables in a logistic regression, where membership in the online or paper group is the dependent variable and the independent variables are various score and demographic variables, such as gender, ethnicity, English language status, status with respect to receiving special education, parents' level of education, etc. The logistic regression model is:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n ,$$

where *P* represents the probability that an examinee with a given set of demographic characteristics (x_1, x_2, …,x_n) is in the online group rather than the paper group. The propensity score is computed using

$$Y = b_1 x_1 + b_2 x_2 + \ldots b_n x_n ,$$

where x_1, x_2, ...x_n represent the examinee's values of the selected demographic variables and b_1, b_2, ...b_n are the estimates of β_1, β_2, ...β_n in the logistic regression model.

Yu et al. (2004) used propensity scoring to investigate differences between computer-based and handwritten essays. In their study, the propensity scores were divided into 20 different intervals, and weights were assigned to the frequencies in each interval for the computer group that reflected the ratio of computer to paper examinees in that interval. Through this process, the weighted sample of computer group closely resembled the paper group in the distribution of the propensity score. Because the propensity score emphasized the demographic variables on which the two groups most differed, the weighted computer group sample was considered demographically similar to the paper group.

In general, MSCA and propensity scoring are used to accomplish the same goal, which is to create statistically matched samples for analysis. Propensity scoring seems as if it might be more attractive when a large number of demographic variables are available rather than just a handful of score variables. MSCA better exploits the covariate relationships between criterion scores and matching score variables than does propensity scoring. For this reason, we believe that MSCA is a better approach than propensity scoring when strong covariates are available. It should be noted that propensity scoring can easily be fit into the bootstrapping approach used with MSCA to draw repeated matched samples, and there may be applications where the use of propensity scoring would make better sense than regression-based MSCA.

Applications of MSCA in K-12 Comparability Studies

Pearson began utilizing MSCA for comparability studies in 2005. During that time, we have refined and expanded our approach to fit specific situations. In this section of the paper, we summarize MSCA results for 46 studies across five states: Texas, Michigan, South Carolina,

Arizona, and Maryland. These studies have addressed a variety of content areas administered in state assessment programs from grade 6 to high school exit exams.

Texas Comparability Studies

The strategy Texas has adopted for introducing online testing is similar to the strategy that many states are using, where online testing is made available to those districts and schools that are willing and able to pursue it. The Texas Education Agency (TEA) began offering optional online testing for certain grades and subjects of the Texas Assessment of Knowledge and Skills (TAKS) beginning in spring 2005. In general, the general TAKS spring administrations have included online versions at grades 7 through 9. In addition, online versions of TAKS Exit Level re-test administrations have also been offered online.

Pearson has adopted MSCA into the equating and scaling procedures for the TAKS subjects and grades offered in dual mode. Specifically, once the paper version of a test form has been equated, we use MSCA to evaluate the comparability of the online and paper test versions. The results of MSCA include alternate raw score to scale score conversion tables, which may be utilized if mode effects are detected and the differences appear to be statistically and practically significant.

Table 1 summarizes the complete list of Texas comparability studies conducted through March 2008. Listed in this table are the administration date, subject, and grade of each study, and the unadjusted summary raw score statistics for the online and paper groups. The next three columns of the table indicate the score variable(s) used to create matched groups and the equated raw score differences between the online and paper groups at the cut defining proficient, and the cut defining advanced. In these columns, a positive difference indicates that the online version of the test was more difficult than the paper version of the test, and vice versa. The last column of the table indicates whether an alternate conversion table was applied to report the online scores.

With regard to the application of the alternate conversion tables (last column of Table 1), three pieces of information are usually evaluated: the bootstrap standard errors of the linking (e.g., whether differences across the scale are within two bootstrap SEs of zero), the magnitude of the raw score differences, and whether or not the alternate conversion table indicates that proficient or advanced performance levels correspond to different raw scores. The standard error of the linking criterion was borrowed from Dorans and Lawrence (1990), who advised: “To

assess equivalence, it is convenient to compute the difference between the equating function and the identity transformation, and to divide this difference by the standard error of equating. If the resultant ratio falls within a bandwidth of plus or minus two, then the equating function is deemed to be within sampling error of the identity function" (p. 247).

In general, psychometricians working on the Texas program and the TEA evaluate these three pieces of information in determining whether or not to apply the alternate conversion tables that result from MSCA. This evaluation also considers the stakes associated with the tests, which are slightly higher for the Exit Level TAKS than for the regular TAKS since passing the Exit Level TAKS is part of Texas high school graduation requirements.

There are several patterns that can be observed in the data presented in Table 1. First, alternate conversion tables have been applied for the online results in most of the studies (28 of 37). Of the nine studies that did not result in alternate tables, seven involved science or social studies tests. Second, the magnitudes of the mode differences at the proficiency cut score ranged from -0.5 to 2.7, with most differences falling around one raw score point. Differences at the advanced cut score have ranged from -1.5 to 1.4, although most differences at that level are less than one raw score point. Third, the results for most of the ELA comparability studies follow an interesting pattern in that differences at the proficient cut score tend to be positive (indicating the online test is more difficult) but differences at the advanced cut score tend to be negative (indicating the online test is easier). We have attributed these results to the structure of the ELA tests, which include 48 multiple-choice items, three short answer constructed response items, each worth three points, and an extended response essay scored on a rubric ranging from one to four. The essay score is weighted four times in the calculation of the ELA total raw score, which ranges from 0 to 73.

Table 1. Summary of TAKS Online vs. Paper Comparability Studies – 2005 through 2008

Admin.	Subject Area	Grade	Unadjusted Online Data			Unadjusted Paper Data			Subject(s) Used to Match Samples	Mode Adjustment at:		Alternate Table?
			N	Mean	Std	N	Mean	Std		Proficient	Advanced	
Spr. 05	Math	8	1,273	32.60	9.27	158,809	32.76	9.78	Gr. 7 Ma, Rd	0.4	0.3	No
Spr. 05	Reading	8	1,840	40.60	7.16	158,282	40.73	7.35	Gr. 7 Ma, Rd	1.3	0.5	Yes
Spr. 05	Social Studies	8	1,449	33.97	7.73	157,809	33.94	8.35	Gr. 7 Ma, Rd	0.4	0.2	No
Sum. 05[b]	ELA	Exit Level	649	37.52	11.1	719	38.24	10.76	Random Groups[d]	1.4	0.5	Yes
Sum. 05[b]	Math	Exit Level	958	26.76	8.92	1,198	27.47	8.81	Random Groups[d]	0.8	0.5	Yes
Sum. 05[b]	Social Studies	Exit Level	355	29.49	11.17	388	29.19	11.13	Random Groups[d]	-0.4	-0.2	No
Sum. 05[b]	Science	Exit Level	1,004	23.70	7.56	1,197	24.17	7.92	Random Groups[d]	0.5	0.2	No
Spr. 06	Math	8	494	32.17	8.34	96,578	34.26	9.41	Composite[c]	2.7	1.4	Yes
Spr. 06	Reading	8	1,080	40.00	6.76	96,516	40.54	7.26	Composite[c]	0.6	0.3	Yes
Spr. 06	Social Studies	8	1,427	34.85	9.39	95,071	35.08	9.08	Composite[c]	1.1	0.4	Yes
Spr. 06	Science	8	918	32.92	9.86	94,941	32.69	9.16	Composite[c]	0.4	0.2	Yes
Spr. 06	Math	9	3,158	32.72	9.72	94,793	33.75	9.89	Composite[c]	1.6	0.9	Yes
Spr. 06	Reading	9	2,882	31.79	5.00	248,045	32.04	4.84	Composite[c]	0.6	0.2	Yes
Sum. 06[b]	ELA	Exit Level	964	44.20	10.67	12,277	44.26	9.82	Composite[c]	-0.1	-0.9	No
Sum. 06[b]	Math	Exit Level	2,416	26.31	6.1	25,728	27.09	6.68	Composite[c]	0.8	0.3	Yes
Sum. 06[b]	Social Studies	Exit Level	587	25.43	6.48	6,018	26.92	7.80	Composite[c]	0.8	0.4	Yes
Sum. 06[b]	Science	Exit Level	2,856	24.45	6.41	29,706	24.82	6.59	Composite[c]	0.4	0.2	No
Fall 06[b]	ELA	Exit Level	1,687	47.28	12.13	22,765	47.81	10.81	Composite[c]	1.6	-1.2	Yes
Fall 06[b]	Math	Exit Level	3,379	28.73	8.22	43,066	29.26	8.26	Composite[c]	1.0	0.4	Yes
Fall 06[b]	Social Studies	Exit Level	1,434	34.41	10.22	16,264	32.60	9.61	Composite[c]	-0.5	-0.3	No
Fall 06[c]	Science	Exit Level	3,705	26.46	7.39	46,821	26.64	7.03	Composite[c]	0.7	0.3	Yes
Spr. 07	Math	7	746	31.25	8.78	111,880	35.12	8.72	Composite[c]	2.2	0.8	Yes
Spr. 07	Reading	7	1,086	37.48	7.02	108,875	39.30	6.29	Composite[c]	1.8	0.6	Yes
Spr. 07	Math	9	3,820	34.16	6.42	101,304	35.42	9.87	Composite[c]	1.6	0.9	Yes
Spr. 07	Reading	9	2,312	32.69	3.91	252,291	32.76	4.28	Composite[c]	1.0	0.6	Yes
Spr. 07	ELA	10	1,313	56.09	8.71	220,046	55.44	8.35	Composite[c]	0.9	-1.1	Yes

Table 1. Summary of TAKS Online vs. Paper Comparability Studies – 2005 through 2008

Admin.	Subject Area	Grade	Unadjusted Online Data			Unadjusted Paper Data			Subject(s) Used to Match Samples	Mode Adjustment at:		Alternate Table?
			N	Mean	Std	N	Mean	Std		Proficient	Advanced	
Sum. 07[b]	ELA	Exit Level	853	43.67	10.95	8,648	42.98	9.84	Composite[c]	0.0	-1.5	Yes
Sum. 07[b]	Math	Exit Level	2,643	25.19	6.5	21,064	26.15	6.88	Composite[c]	1.2	0.5	Yes
Sum. 07[b]	Social Studies	Exit Level	591	26.36	6.08	5,446	26.39	6.50	Composite[c]	0.2	0.1	No
Sum. 07[b]	Science	Exit Level	3,065	25.84	9.84	25,015	26.03	6.13	Composite[c]	0.5	0.2	No
Fall 07[b]	ELA	Exit Level	2,257	48.00	11.76	19,271	47.71	10.76	Composite[c]	1.2	-0.9	Yes
Fall 07[b]	Math	Exit Level	4,497	29.13	8.64	27,954	29.80	8.62	Composite[c]	1.3	0.5	Yes
Fall 07[b]	Social Studies	Exit Level	1,968	33.56	10.27	15,883	33.25	9.76	Composite[c]	0.5	0.2	Yes
Fall 07[b]	Science	Exit Level	4,745	27.48	7.69	39,456	27.74	7.42	Composite[c]	0.9	0.4	Yes
Spr. 08[a,b]	Math	Exit Level	1,822	28.54	8.07	24,841	29.04	8.06	Composite[c]	1.2	0.5	Yes
Spr. 08[a,b]	Social Studies	Exit Level	541	32.67	10.41	6,900	31.89	9.62	Composite[c]	0.8	0.4	Yes
Spr. 08[a,b]	Science	Exit Level	1,694	26.53	7.33	23,430	27.05	7.23	Composite[c]	1.1	0.5	Yes

[a] The assessment was administered in March 2008.

[b] For retest administrations.

[c] Composite was based on their scores from the previous year's primary TAKS administrations on ELA (or reading), math, science and social studies. For exit level retest students, the composite was based on their scores on the exit level primary TAKS administrations.

[d] Summer 2005 exit level comparability studies involved randomly assigning students to conditions. Matching was not employed

Figure 1 presents the differences between the equated online raw scores and the paper raw scores from the fall 2006 ELA administration. The graph includes the intervals around zero defined by plus and minus two bootstrap standard errors of equating. In this plot, positive differences indicate that the online version of the test is more difficult than the paper test, and vice versa.

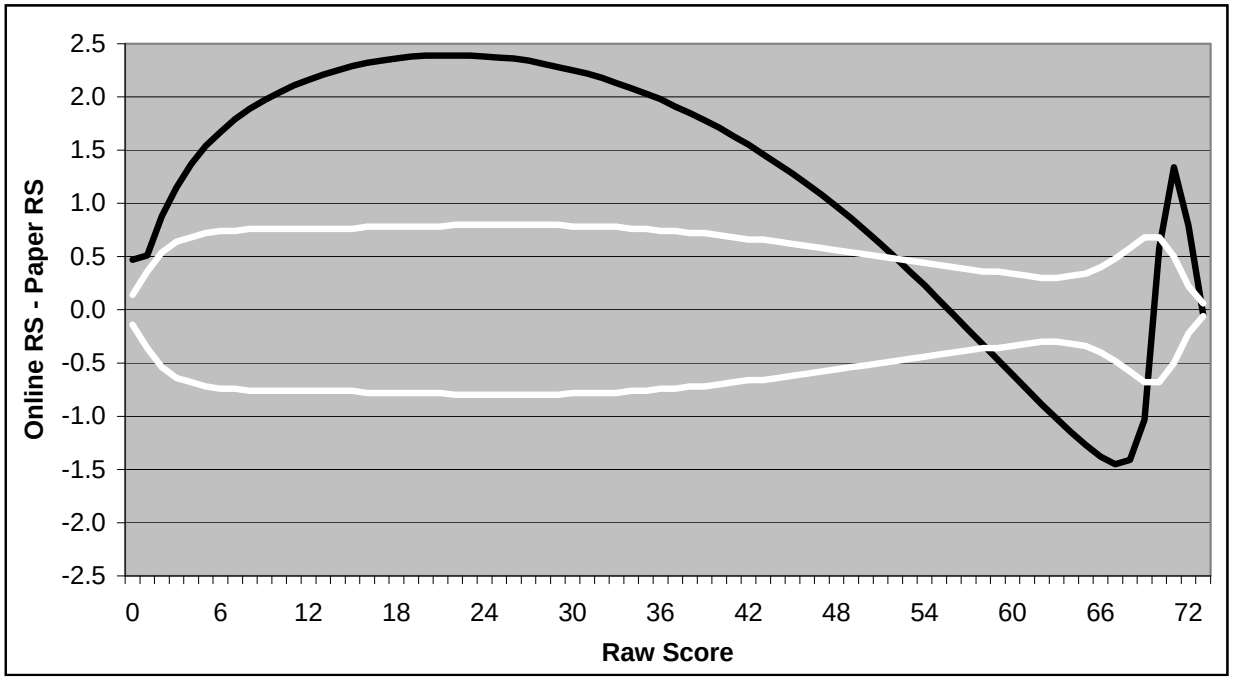

Figure 1. TAKS ELA Online Minus Paper Raw Score Differences and ±2 Equating Standard Errors

In Figure 1, the proficient raw score cut on the online test was 41 and the advanced raw score cut was 64. The mode differences seen in Figure 1 are a result of an interaction between the multiple-choice and the extended essay portion of the ELA test and are typical of consistent results we have seen over time. The multiple-choice items tend to measure better at the low raw score levels. For these items, online performance tends to be lower than paper performance. The short-answer items are difficult but there tends not to be clear evidence of mode effects in either direction. The essay portion of the test also tends to be difficult and online performance tends to be higher than paper performance. Because the essays carry a heavy weight in the overall raw score, the equating differences become negative at the higher raw score levels. The alternate score conversion tables resulting from the MSCA serve to correct the mode effects detected in both directions.

Michigan Comparability Studies

As part of the fall 2005 administration of the Michigan Educational Assessment Program (MEAP), the Michigan Office of Educational Assessment and Accountability (OEAA) undertook a study of tests delivered by computer for grade 6 ELA and social studies tests. The study drew primarily from schools that had been provided with laptop computers as part of a statewide initiative called Freedom to Learn (http://www.ftlwireless.org/). Approximately 1,100 students across 19 school districts participated in the study. For the participating schools, online tests were administered to grade 6 students using the laptop computers that were assigned to them. In addition, a small number of schools participated by administering the online MEAP tests through computer facilities maintained in classrooms or computer labs.

Participating schools tested online during the same administration window as the paper-and-pencil MEAP. Provisional (raw) scores on the online tests were provided within 48 hours of the online administration. These included multiple-choice total scores and scores on grade-level expectations. In addition, preliminary scores on constructed response (CR) items for the online tests were generated using an automated essay scoring engine, referred to as the Intelligent Essay Assessor (IEA; Foltz, Latham, & Landauer, 1999). However, CR items were also scored along with responses based on the paper test administration using traditional essay scoring procedures prior to final score reporting.[1]

The comparability data were analyzed using the MSCA approach. In this design, test scores from the spring 2005 administration of the grade 5 social studies test, gender, and ethnicity were used as matching variables. Since grade 5 students did not take ELA in spring 2005, the grade 5 social studies scores were as a matching variable for both the social studies and ELA test comparisons. Separate analyses were done for students for the reading and writing portions of the ELA test, as these areas are scaled and reported separately. For the CR questions, the scores used for the comparability analyses were the final reported scores. As directed by the OEAA, the reported CR scores for online students were based on the *higher of the automated and human ratings*.

Table 2 presents summary statistics (mean, standard deviation, minimum and maximum scores) for the social studies, reading, and writing scores for the grade 6 students included in the

[1] The social studies test had one three-point CR item, the reading test had one six-point CR item, and the writing test had one six-point CR item and one four-point CR item.

comparability study. There were 1,095 online students and 112,729 paper students in the social studies comparability samples. For ELA, there were 1,133 online students and 42,872 students in the paper sample. The ELA statistics in Table 2 are presented separately for reading and writing, as these two measures are separately scaled on the operational test. There are two entries of summary statistics for the online tests. One entry is based on using only the human scores for the CR items. The second entry is based on using the higher of the automated and human scores for the CR items. As would be expected, mean raw scores are higher when the CR scores are based on the higher of the automated and human scores.

Table 2. Fall 2005 Raw Scores and Spring 2005 Social Studies Scale Scores

	Grade 6 Social Studies RS					Grade 5 Social Studies Scale Score				Corr.
Data Set	N	Mean	Std	Min	Max	Mean	Std	Min	Max	Gr5-Gr6
Online Sample[1]	1095	28.96	8.64	7	48	506.40	37.42	399	620	0.79
Online Sample[2]	1095	29.16	8.57	8	48	506.40	37.42	399	620	0.79
Paper Sample	112729	29.60	9.35	1	49	505.96	40.74	296	748	0.76
	Grade 6 Reading/Writing RS					Grade 5 Social Studies Scale Score				Corr.
Data Set	N	Mean	Std	Min	Max	Mean	Std	Min	Max	Gr5-Gr6
Online Reading[1]	1133	26.12	6.80	5	40	507.19	37.02	399	620	0.69
Online Reading[2]	1133	26.29	6.80	6	40	507.19	37.02	399	620	0.69
Online Writing[1]	1133	6.72	2.02	0	14	507.19	37.02	399	620	0.49
Online Writing[2]	1133	7.45	2.09	0	14	507.19	37.02	399	620	0.45
Paper Reading	42872	26.47	7.40	0	43	505.10	41.11	296	748	0.71
Paper Writing	42872	7.11	1.97	0	15	505.10	41.11	296	748	0.51

[1]Based on human constructed response scores
[2]Based on the higher of human and automated constructed response scores

Figure 2 presents differences between the online and paper MEAP score conversions along with the intervals defined by plus and minus two bootstrap standard errors of equating. In these plots, positive scale score differences indicate that the online version of the test is more difficult than the paper test, and vice versa. It be seen that for social studies and reading, the differences are close to or exceed the +2 SE interval over most of the score range. For writing, the differences are close to or exceed the -2 SE interval over most of the score range.

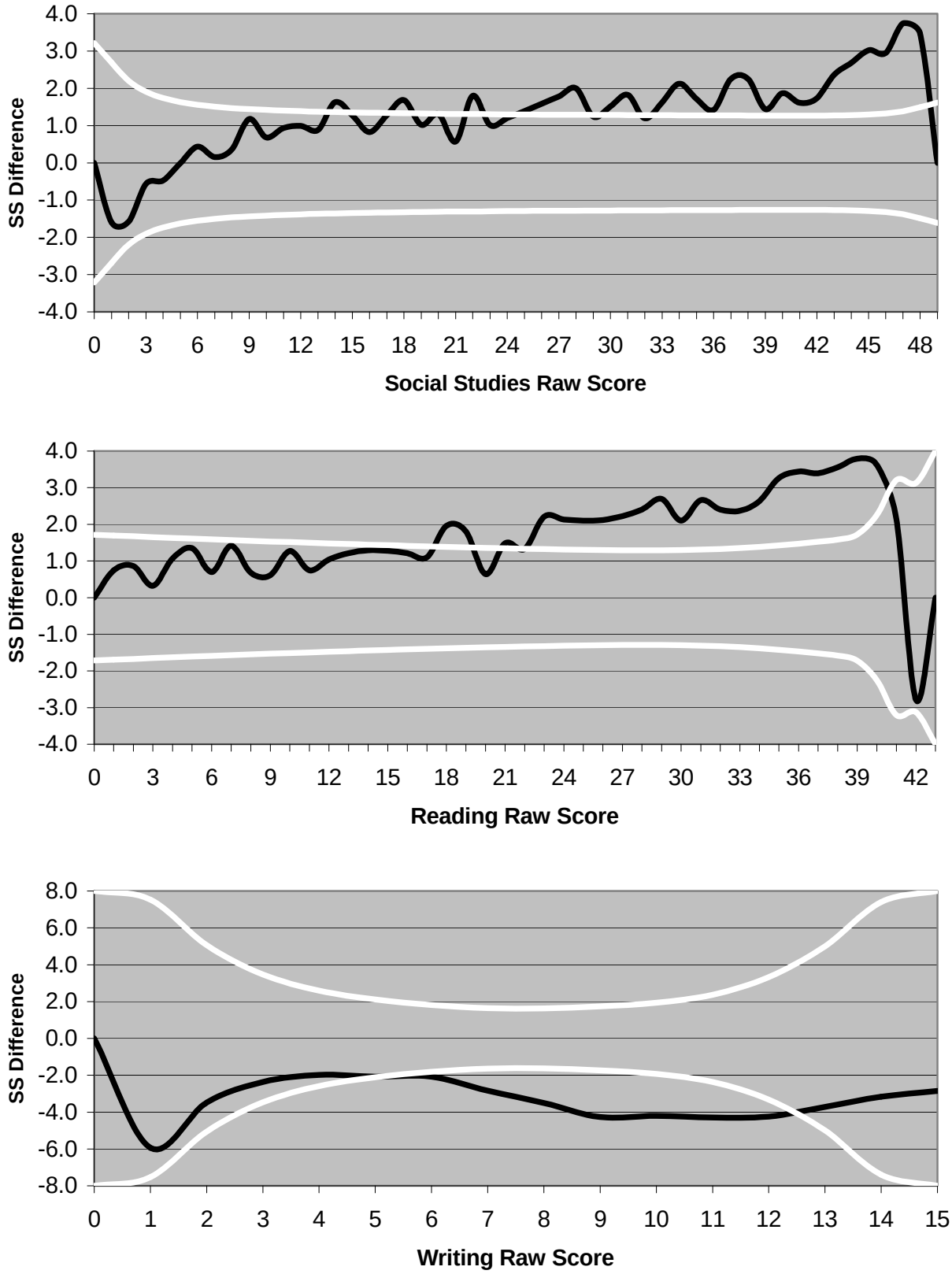

Figure 2. MEAP Online Minus Paper Scale Score Differences and ±2 Equating Standard Errors (Online Scores based on Higher of Human and Automated CR Ratings of CR Items)

Based on the bootstrap equating results OEAA approved using the paper raw to scale score conversion table for the online administrations of the social studies, reading, and writing tests. The rationale for this decision was as follows:

1) the "met standard" cut scores for the online and paper conversions corresponded to the same raw score for all three tests
2) The online versus paper scale score differences were mostly within the ±2 SE interval, especially in the vicinity of the "met standard" cuts
3) Mode differences for reading and writing tests cancelled each other out when combined to create the English language composite score

It was clear from the data that the policy decision to report the higher of the automated and human scores on the constructed response items affected the comparability study results, especially for the writing test. When results were compared using only the human scores, the online version of the writing test was found to be significantly more difficult than the paper version.

South Carolina Comparability Studies

The South Carolina Department of Education has conducted investigations of mode comparability as they have implemented optional online testing for their end-of-course examination program (EOCEP). An initial study that did not use MSCA was conducted by Kim and Huynh (2007). A second study by Kim and Huynh (in press) also applied different analysis methodologies to one of the data sets that we analyzed with the MSCA method. Our studies were conducted on the English I exam administered in fall 2005 and the English I and Physical Science exams administered in spring 2006. These multiple-choice tests are administered primarily to grade 9 students. In both studies, grade 8 scores on the Palmetto Achievement Challenge Test (PACT) were utilized with the MSCA. Table 3 summarizes the performance of the paper and online groups used in the study. The data in Table 3 indicate a very strong relationship (i.e., correlations above 0.75) between the previous year's PACT scores and the EOCEP scores for the matched students, which represented an ideal situation for applying the MSCA method.

Table 3. Summary Statistics for Paper and Online Groups Used in South Carolina Mode Comparability Studies

Fall 2005 Administration							
Sample	Test	N	Mean SS	SD SS	Min SS	Max SS	Corr[a]
Paper	EOCEP English I	3169	76.4	11.0	44	100	0.773
Group	PACT G8 ELA	3169	805.9	12.0	747	845	
Online	EOCEP English I	768	78.7	10.9	37	100	0.771
Group	PACT G8 ELA	768	808.2	12.1	774	845	
Spring 2006 Administration[b]							
Sample	Test	N	Mean	SD	Min	Max	Corr[a]
Paper	EOCEP English I	30,703	35.6	8.8	1	53	0.762
	PACT G8 ELA	30,703	804.0	12.6	744	848	
Online	EOCEP English I	4,160	36.3	8.4	9	53	0.767
	PACT G8 ELA	4,160	805.6	12.6	748	848	
Paper	EOCEP Physical Science	26,131	29.1	9.2	1	55	0.774
	PACT G8 Science	26,131	805.2	17.3	736	864	
Online	EOCEP Physical Science	3,893	30.1	9.0	3	55	0.750
	PACT G8 Science	3,893	806.6	17.0	750	864	

[a] Corr refers to the correlation between the relevant PACT and EOCEP tests
[b] EOCEP scores reported for spring 2006 are raw scores rather than scale scores.

The results of the comparability analyses suggested no mode effects between the online and paper versions of the EOCEP English test based on either the fall 2005 or spring 2006 administrations. Curiously, the EOCEP Physical Science results suggested higher performance for the online group compared with the paper group. Figure 3 presents differences between the online and paper EOCEP Physical Science score conversions along with the intervals defined by plus and minus two bootstrap standard errors of equating. In this graph, the differences are characterized by jagged discontinuities that are due to the fact that the differences and bootstrap standard errors were based on integer scale scores obtained by table lookup rather than unrounded scale scores (which are typically used in the calculations). Although the scale score differences fell outside of two bootstrap standard errors, they were less than one scale score point at score levels beyond chance level. On the raw score scale, the differences were less than one-half of a point. Because these differences did not seem to be of practical significance and because no hypotheses to explain higher online performance were offered, the differences found for the EOCEP Physical Science were not judged to warrant further investigation.

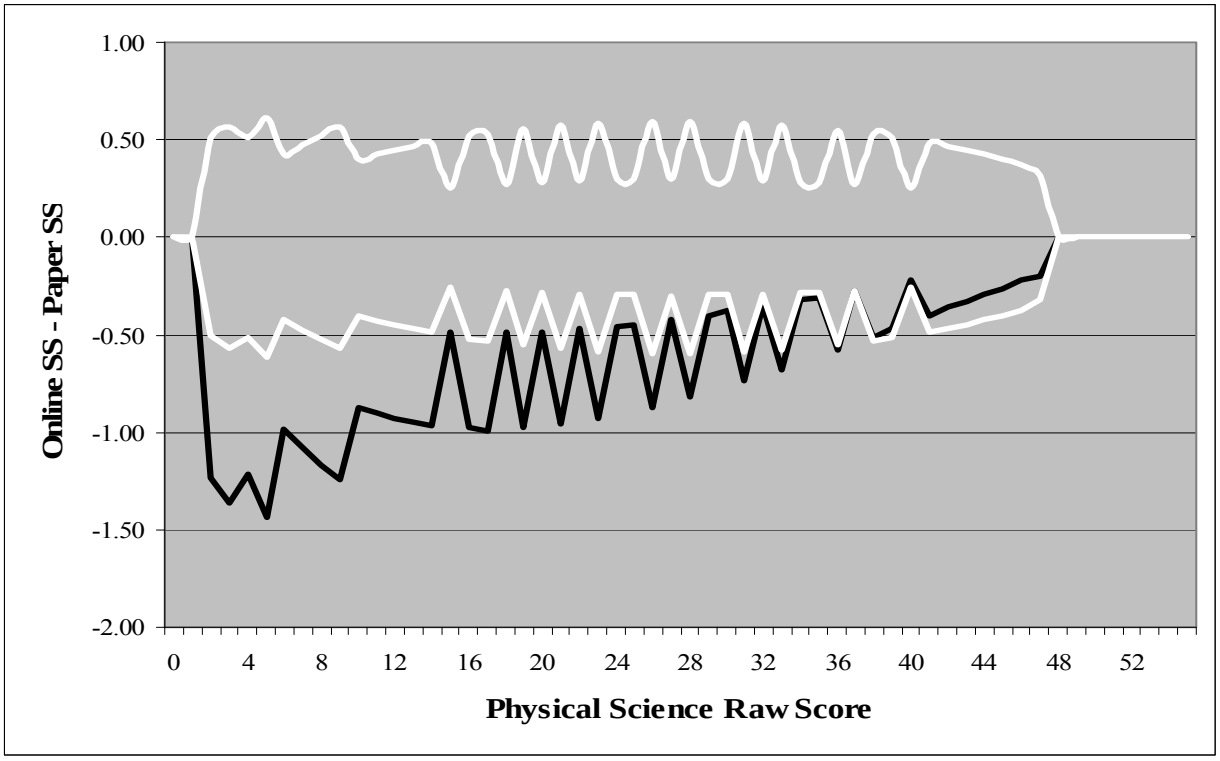

Figure 3. EOCEP Physical Science Online Minus Paper Scale Score Differences and ±2 Equating Standard Errors

Comparability Study for Arizona Grade 8 Science[2]

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.

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,

[2] The complete report of this study is available from the Arizona Department of Education web site at: http://www.azed.gov/standards/aims/Administering/aims_science_comparability_report_revised.pdf

mathematics, and writing AIMS tests. In addition, we used gender, ethnicity[3], and the field-test form administered as matching variables for the study.

For these comparability analyses, a variation on the MSCA was used. Specifically, the study was carried out using the three-parameter logistic (3PL) model and the MULTILOG program version 7.0.3 (Thissen, Chen & Bock. 2003). In addition, the focus for comparisons was IRT ability parameter (theta) estimates rather than raw scores and scale scores. This approach was utilized because field-test rather than operational data were used for the MSCA analyses. 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. These theta estimates were used as the basis for comparisons based on the MSCA.

Table 4 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 4 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.

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

Table 4. Summary Statistics for AIMS Science Field-Test Theta Estimates and Raw Scores

Form	N	Theta				Raw Score			
		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

To implement the matched samples approach, we considered the operational AIMS scale scores in reading, mathematics, and writing, gender, ethnicity, and the assigned field-test form as matching variables for the comparability analyses. To simplify the matching process, we used multiple regression procedures to produce a composite score variable that was grouped into 15 equally-sized intervals. The regression model used to create the composite score was as follows:

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

Within each of 100 bootstrap iterations, 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 composite score group, gender, ethnic group, and field-test form, also with replacement.

The bootstrap approach permitted us to compare the paper and online groups within iteration and also to aggregate the results over iterations. We compared the groups in terms of mean standardized differences between theta estimates and mean effect sizes.

Table 5 presents the mean theta estimates and mean differences in theta estimates for the online and matched paper groups across 100 bootstrap replications. These results are presented for all students and separately by gender and ethnic groups across the five field test forms

combined. Results include mean differences, standard deviations of the differences, standardized differences, and effect sizes. 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 5 lists the "average N" over replications.

Table 5. Summary of AIMS Grade 8 Science Estimated Theta Differences Based on the Matched Sampling Results

Group	Ave. 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

Overall, the results in Table 5 suggest comparable AIMS 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. Similarly small differences were also found by gender and for different ethnic groups. From these results, the field-test performances of the paper and online samples were judged sufficiently comparable to support equating and reporting scores for the AIMS operational grade 8 science test without regard to testing mode.

Maryland Grade 5 and 8 Science Comparability Studies

The Maryland MSA Science field-test for grades 5 and 8 was conducted in spring 2007 and included both a traditional paper-and pencil administration and a computer-administered, online format. The comparability study data included both multiple-choice (MC) items and brief constructed-response (BCR) item responses (scored 0-2) for each participating grade 5 and 8 student on one of 10 randomly administered field-test forms[4]. The matching variables included

[4] The same science forms were randomly spiraled for students testing by paper and randomly assigned by the test delivery system to students testing online.

scale scores on the spring 2007 operational MSA reading and mathematics tests administered by paper only. In addition, we matched students on gender, ethnicity, and the field-test form that was administered to them (e.g., students taking form 1 online were matched to students taking form 1 on paper, etc.).

The comparability analyses utilized the variation on the MSCA approach as described for in the previous section of this paper. We first calibrated the online science field-test data using MULTILOG, obtaining item and ability parameter (theta) estimates. Next, we did a second run of MULTILOG with the paper field-test data, fixing the item parameters at the values obtained from the online 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 6 summarizes the numbers of students participating in the field test in paper and online formats and presents descriptive statistics for the estimated thetas, raw scores, and scale scores used in matching samples. As the data in Table 6 indicate, the online group had higher mean estimated thetas, higher mean raw scores, and higher means on the matching scale scores compared with the paper group at both grades 5 and 8.

Table 6. Summary Statistics for Science Field-Test Theta Estimates, Raw Scores and Matching Scale Scores

		Theta				Raw Score			
Form	**N**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Grade 5 Online	20,527	0.02	1.04	-4.00	4.00	46.23	12.16	0	80
Grade 5 Paper	39,831	-0.30	1.10	-4.00	4.00	42.48	12.53	0	82
Grade 8 Online	24,256	0.02	1.04	-4.00	3.77	43.72	13.66	0	80
Grade 8 Paper	38,359	-0.34	1.16	-4.00	3.83	39.39	14.53	0	81
		Reading				**Math**			
Form	**N**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Grade 5 Online	20,527	416.37	34.80	240	598	428.33	36.01	240	576
Grade 5 Paper	39,831	409.09	36.38	240	619	419.90	38.07	240	576
Grade 8 Online	24,256	411.18	29.58	240	567	425.09	36.62	240	586
Grade 8 Paper	38,359	402.73	32.97	240	567	414.60	39.51	240	591

Table 7 presents the mean theta estimates for the online and matched paper groups and the mean estimated theta differences between the two groups across 100 bootstrap replications. These results are presented for all students and separately by gender and ethnic groups across the 10 field test forms combined. 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 7 lists the "average N" over replications.

Table 7. Summary of Estimated Theta Differences Based on the Matched Sampling Results for MSA Grade 5 and Grade 8 Science Field-Tests

Group	Ave. N	Mean Theta		Mean Diff.	SD of Diff.	Standardized Difference	Effect Size
		Online	Paper				
All Grade 5 Field-Test Forms Combined							
All Students	3000	0.01	-0.05	0.06	0.01	4.31*	0.06
White	1192.12	0.28	0.22	0.07	0.02	3.48*	0.07
Black	1323.38	-0.61	-0.64	0.03	0.03	0.88	0.03
Others	484.5	0.04	-0.09	0.13	0.05	2.46*	0.13
Female	1493.49	-0.03	-0.09	0.06	0.02	2.91*	0.06
Male	1506.51	0.06	0	0.07	0.02	3.00*	0.06
All Grade 8 Field-Test Forms Combined							
All Students	3000	0.02	-0.02	0.03	0.02	2.04*	0.03
White	1888.57	0.31	0.29	0.02	0.02	0.84	0.02
Black	877.8	-0.62	-0.68	0.06	0.03	1.89	0.06
Others	233.63	0.05	-0.02	0.07	0.05	1.34	0.06
Female	1510.18	-0.01	-0.05	0.03	0.02	1.56	0.03
Male	1489.82	0.05	0.01	0.03	0.02	1.49	0.03

Overall, the MSCA results in Table 7 suggests that the online group performance on the MSA grade 8 science field-tests was higher than the paper group performance for both grades 5 and 8. The mean difference in theta estimates was 0.06 for grade 5 and 0.03 for grade 8; the standardized difference was 4.31 for grade 5 and 2.04, for grade 8. Although the standardized differences were greater than 1.96, the effect size was 0.06 for grade 5 and 0.03 for grade 8.

These effect sizes suggest that the differences are not very meaningful from a practical standpoint.

In following up the MSCA results by comparing performances of individual items across modes, we found that in both grades 5 and 8, all BCR items that were flagged for "significant" mode differences were easier in the online format than they were in the paper format. There were six of 40 BCR items flagged for mode effects at grade 5 and 16 of 40 BCR items flagged at grade 8. Although a non-speculative explanation for this finding was not apparent, it was recommended that the online and paper performance on the BCR items be closely monitored as the grade 8 science tests are implemented operationally.

Discussion of K-12 Comparability Study Results

Generalizing across the studies presented in this paper, several trends seem to emerge. First, the studies suggest that performance on paper and online versions of traditional multiple-choice science tests is likely to be comparable. Although some of the Texas studies involving science resulted in the use of alternative conversion tables, raw score differences at the proficient and advanced cut scores were consistently less than one point. Furthermore, the studies in South Carolina, Arizona, and Maryland all suggested that students taking science tests online in these states would not be disadvantaged. In fact, the comparability study results in South Carolina and Maryland suggested slightly higher performance for students taking science tests by computer. Comparability results for other subjects seem less clear. The Texas MSCA studies indicated that social studies results were similar to those for science, sometimes indicating slight mode effects and sometimes not. In Michigan, results of the grade 6 comparability study for social studies indicated slight mode effects, but did not suggest the need for an alternate conversion table. However, the constructed response item in the social studies test was more difficult online than it was on paper. For Mathematics, the only comparability studies we reported were from Texas, and most of these indicated mode effects of a raw score point or more.

Results for Reading/ELA are more complicated to interpret. In Texas, reading comparability studies consistently indicated mode effects. For ELA, online performance in Texas was consistently higher than paper performance on the essay portion of the test, with the exception of the first study done in summer 2005. However, results over studies have also consistently indicated that the rest of the ELA test is more difficult in the online format compared with the paper format. In South Carolina, two studies each found no evidence of mode

effects for the multiple-choice EOCEP English I test. In Michigan, there was evidence of slight mode effects for the grade 6 reading test, but when final essay scores were based on human scoring rather than the higher of human and automated scores, the writing test was clearly more difficult for students testing online than it was for students testing on paper. These conflicting results involving English tests can be attributed to the different grade levels of the test takers and especially the differences in the formats of the tests used in Texas, South Carolina, and Michigan.

A final observation from reviewing the results of these many studies across five states concerns the application of MSCA. Because the method uses score covariates and other demographic variables to adjust for differences in the online and paper groups, the initial unadjusted performance of the online and paper groups should not be related to the performances of the groups after MSCA have been applied. However, looking closely at the data presented in this paper, one will see that there seems to be a relationship between the unadjusted means of the online and paper groups, and the detection and direction of mode effects based on the MSCA results. We have done sensitivity analyses and simulation studies that suggest MSCA is relatively robust in the presence of mode effects that are confounded with the ability levels of the online and paper groups (Way, Davis & Fitzpatrick, 2006; Way, Um, Lin & McClarty, 2007). Nevertheless, the pattern that is apparent across the studies summarized in this paper serves to remind us that the effects we are trying to detect are small and subtle, and that methods like MSCA are still rather blunt instruments for the task at hand.

Online versus Paper Comparability Studies – Looking to the Future

The studies summarized in this paper are sort of a microcosm of the general literature on comparability studies. First, results of our studies have been mixed. As with the general literature, some of our studies have reported mode effects and others have not. Second, when effects have been detected, they have typically been small, usually in the neighborhood of a half point to a point and a half on the raw score scale. Third, while it seems to have been worth it to apply MSCA in these studies, there remains the doubt associated with the use of a quasi-experimental method. Yet, such methods seem to be the only choice for analysis, and there is no reason to expect strong and tightly-controlled experimental designs will be any more feasible in the future than they have been up until now.

One question that might be asked at this point is how long should test developers expect to continue conducting comparability studies? In Texas, for example, comparability studies have become woven into the equating process for those TAKS tests that are offered in both modes. This seems to be a satisfactory approach for the near term but could become untenable in the future for any number of reasons. In South Carolina, based on their comparability study results, online and paper versions of EOCEP tests are offered side by side and scores are assumed to be equivalent. While South Carolina may have reason to monitor the online and paper performance over time, the state does not consider comparability to be an issue for the EOCEP program. Similar logic may be used by Arizona and Maryland with respect to their science tests, based on the MCSA results reported in this paper.

The different comparability results across different states are troubling. In compiling results of the studies included in this paper, we have been forced to wonder if there is something in the water down in Texas that makes it more difficult for kids to take tests on computer. Obviously, we would like a more rational basis for sorting out these mode effects.

Theoretical Considerations of Mode Effects

One helpful consideration in interpreting studies evaluating the comparability of paper and online versions of the same tests is theoretical explanations as to why different test items perform differently across modes. Some explanations seem intuitive: reading items that require scrolling test are likely to be more difficult for students who are used to reading from books and paper. Or, mathematics items that require examining a graph or geometric figure may be more difficult to process when presented on computer than on paper. These explanations suggest that we are measuring the same construct, but the online format is making it more difficult for students to perform their best on the item. However, these explanations should also lead to verifiable predictions about which items in a test are likely to demonstrate mode effects. Thus, one might expect items associated with longer reading passages to be more relatively more difficult for students testing online than items associated with shorter passages. Similarly, those mathematics items involving more complex graphs or figures should be the ones flagged as being more difficult for students testing online. If theoretical explanations for mode effects can be empirically verified over time, it becomes easier to explain them and take actions to address them as paper and online tests continue to be given side by side. On the other hand, if no

consistently understandable explanation can be offered for mode effects over time, it becomes tempting to question whether the effects being detected are truly real. We have done some follow up work of mode effects on individual items as part of the TAKS program (Keng, McClarty & Davis, 2006), but more work along these lines is needed.

A more challenging example is extended writing tasks. It seems reasonable to assume that students who engage in word processing in their everyday school work will perform better when they take a writing test online. On the other hand, students with little exposure to word processing in their daily instruction are less likely to perform as well on a computer-administered extended writing task. In this case, not only is there a construct equivalence question related to handwriting versus word processing, there is also an interaction of instruction and assessment. In this case, assessing students in a manner that is consistent with how they are instructed may be more important than establishing comparability between a paper and online versions of the test. Furthermore, research evaluating computer interfaces and how students are approaching writing tasks should become part of assessing comparability (Kong, Lin, Strain-Seymour & Davis, 2008).

Comparability Analyses as DIF Detection

When the same items are administered in both online and paper modes, it is relatively straightforward to apply statistical techniques for detecting differential item functioning (DIF). Such analyses can be used in linking online and paper versions of the same exam in a manner similar to the way that analyses to evaluate equating item sets prior to test equating are used. In equating settings, the interest is usually in identifying anchor items that have changed significantly in performance from their previous use compared with the other items in the set. In the case of linking online and paper versions of a test, the interest would be in identifying items that do not perform the same in the paper and online formats. As in equating analyses, these items could be eliminated from the common item set. In this case, the analyses used to develop score conversions would treat the item in the online and paper formats as two unique items.

The idea of using DIF to address mode effects makes the most sense in situations where one believes that only a handful of items might be characterized by performance differences. DIF analyses cannot be used to correct for mode effects that influence all items on a test equally. In this case, the mode effects represent differential test functioning rather than differential item functioning. Because the idea of addressing mode effects through DIF techniques is both

compelling and feasible, it is critically important to develop theories of how mode effects are manifested in online versions of tests. Without better theoretical guidance, the options for addressing mode effects over time are not very attractive: one can declare victory and deny that they are of practical significance, one can resist offering tests online at all, or one can be resigned to indefinitely applying sophisticated statistical techniques that will never be able to fully overcome the weak conditions under which the data for the comparability studies are collected.

Online Testing – Buyer Beware

One analogy for thinking about comparability between online and paper tests in K-12 settings is the way that popular drugs are advertised. We all have seen commercials for new drugs that are supposed to help you sleep better or lower your blood pressure or help you be less anxious. These drugs typically have known potential side effects. The drug companies are required to communicate these side effects to consumers and consumers are urged to consult with their doctors in considering whether or not they want to try the drug. Online versions of tests might be thought of in a similar manner. That is, potential users of online tests are urged to consider whether this mode of testing is right for their school and their students. With the possible exception of writing, most experts would agree that the same tests offered in online and paper versions are measuring the same constructs. Furthermore, mode effects in comparability studies have consistently been found to be small or nonexistent. Finally, surveys indicate that students testing online enjoy taking tests by computer, and tend to prefer it to traditional paper testing (Glassnapp, Poggio, Poggio, & Yang, 2005; O'Malley et al., 2005; Ito & Sykes, 2004). There are clearly benefits to online testing, but there could be side effects: if students are not comfortable using computers, if they are not used to reading text online or using word processing tools to complete classroom assignments, online testing might not be for them. Informed school personnel can make informed individual decisions with regard to online testing. Within this "buyer beware" logic, it seems appropriate to hold all test-takers to the same performance standards regardless of the mode they tested in, especially since a tightly controlled design for evaluating mode comparability is typically not feasible.

Conclusions

The purpose of this paper was to summarize a number of studies that we have conducted with K-12 state departments of education using a particular analysis method referred to as Matched Samples Comparability Analyses (MCSA). Similar to the comparability study literature

in recent years, the results of our studies have been mixed. In some cases, analyses have indicated mode effects and in others, no mode effects of practical significance were detected. We discussed trends across the studies and reached some tentative conclusions about those subject areas where mode effects seem to be less of a concern (science, social studies) and subject areas where mode effects seem more likely to be found (reading/ELA, mathematics).

Looking towards the future, the difficulties of collecting appropriate and reliable comparability data and the general status of K-12 online testing raise some doubts about the value of comparability studies. The current state of K-12 testing is dominated by the assessment model that NCLB has imposed upon the states, and it is particularly difficult to transition tests online within this model. The need for comparability studies stems from the emphasis on score equivalence in the *testing standards*, but such studies do not seem to serve the transition to online testing well. K-12 online testing has reached a crossroads. Full transitions to online testing have been rare, and many testing experts are predicting that it will be another 10 or 15 years before paper-based testing is completely replaced with online testing. The reasons are complex, but at least some of them are related to the traditional summative testing model that has been embraced in the NCLB legislation. Before state departments, school districts and students are able to fully experience the benefits of computer-based assessment, it is likely that our paradigm of K-12 testing will have to shift. The emerging emphasis on formative and interim assessments that are linked to or even embedded within instructional systems provides one indicator for the future of online testing. As this future is realized, it is possible that the concept of score equivalence will take on completely new and different meanings.

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). Standards for educational and psychological testing. Washington, DC: AERA.

Bennett, R.E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment, 1(1)*. Available from http://www.jtla.org.

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

Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test forms. Applied Measurement in Education, 3, 245-254.

Foltz, P. W., Latham, D., & Landauer, T. K. (1999, October). The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1(2). Retrieved October 23, 2005, from http://imej.wfu.edu/articles/1999/2/04/index.asp.

Glasnapp, D.R., Poggio, J., Poggio, A., & Yang, X. (2005). Student attitudes and perceptions regarding computerized testing and the relationship to performance in large scale assessment programs. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, CA.

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.

Kim, D. H., & Huynh, H. (2007). Comparability of computer-based and paper-and-pencil testing for algebra and biology examinations. The Journal of Technology, Learning, and Assessment, 6(5), 1-30.

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

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.

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

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

Kong, J., Lin, C., Strain-Seymour, E., & Davis, L.L. (2008). Evaluating the comparability between online and paper assessments of essay writing. Presented at the annual meeting of the Association of Test Publishers, Dallas, TX.

McDonald, M.E. (2007). Preparing students for the licensure exam: The importance of the NCLEX. In McDonald, M.E., The Nurse Educator's Guide to Assessing Learning Outcomes (2nd Edition). Sudbury, MA: Jones and Bartlett Publishers, Inc.

Mayfield, K. (2002). A D-minus for computer exams. Wired. Retrieved February 15, 2008 from http://www.wired.com/culture/education/news/2002/08/54459.

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.

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.

Rosenbaum. P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician, 39(1), 33–38.

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.

Wang, S., Jiao, H., Young, M., Brooks, T., & Olsen, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests.Educational and Psychological Measurement. 67(2), 219-238.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olsen, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments. Educational and Psychological Measurement, 68(1), 5-24.

Way, W. D., Um, K., Lin, C., & McClarty, K. (2007). 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). 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. Available at http://www.pearsonedmeasurement.com/downloads/research/RR_06_01.pdf.

Yu, L., Livingston, S.A., Larkin, K.C., & Bonet, J. (2004). Investigating differences in examinee performance between computer-based and handwritten essays (ETS Research Report RR-04-18). Princeton, NJ: Educational Testing Service.