

Summary of the Online Comparability Studies for One State's End of Course Program

Susan M. Lottridge, Ph.D.
W. Alan Nicewander, Ph.D.
Howard C. Mitzel, Ph.D.

This report was based on research funded by an Enhanced Assessment Grant from the US Department of Education, awarded to the North Carolina Department of Public Instruction, in partnership with the Council of Chief State School Officers and its SCASS TILSA and SCASS CAS. Publication of this document shall not be construed as endorsement of the views expressed in it by the US Department of Education, the North Carolina Department of Public Instruction, or the Council of Chief State School Officers.

June 5, 2009



Pacific Metrics Corporation
585 Cannery Row, Suite 201
Monterey, California 93940

Acknowledgements: The authors thank the research and state members of the Enhanced Assessment Grant for their guidance and feedback on this study. In particular, the authors thank Phoebe Winter for her continued support and direction, and Laura Kramer for working so closely with them and providing detailed feedback throughout this process.

Summary of Online Comparability Studies for One State's End of Course Program

Introduction

This report presents a summary of online comparability studies conducted on behalf of North Carolina's Enhanced Assessment Grant. The primary focus of this summary and of the grant was on methods for evaluating comparability of test variations. Thus, this summary will focus primarily on the methodological approaches and issues encountered in the comparability studies. In particular, the use of a new method (propensity score matching) for conducting the comparability studies was evaluated as a possible alternative to a within-subjects design. Additionally, the results are presented in light of the methodologies used.

Five subject areas of one state's End-of-Course program were examined in a series of comparability studies. The subject areas were Algebra I, English I, Biology, Civics & Economics, and U.S History. Courses were taken mostly by high school students, although 8th grade students took the Algebra I course. The reports for these studies are available and the references are listed in Table 1. As stated in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2004) comparability studies must be conducted if tests are given in different modes, such as on paper and online.

Scores from the End-of-Course tests are used to provide feedback to students, teachers, parents, and administrators on what students' know and can do relative to the achievement level standards and relative to other students taking the test. Scores are reported in terms of scale score and whether the student meets one of four hierarchical achievement level standards. The test scores count a minimum of 25% of a student's grade in the course. In addition, students must score as proficient in a series of End-of-Course tests to satisfy the state's high school exit standards.

Method

Two types of designs were used to evaluate the comparability of scores from the online and paper versions of the End of Course tests.

1. Within-subjects design. Each study used a within-subjects design. In the within-subjects design, examinees took both an online and paper version of the test. Counterbalancing was designed into the study to control for order effects—schools were assigned to a test order (online first or paper first). For two subject areas (Algebra I and English I), counterbalancing worked successfully; the number of examinees was evenly distributed across test order conditions and the samples appeared to come from the same population. For three subject areas (Biology, Civics & Economics, and U.S. History), counterbalancing was not implemented in practice. For these subjects, the number of examinees was not evenly distributed across test order conditions, presumably due to school self-selection into test order conditions.

2. Between-subjects design. Between-subjects follow-up studies using propensity score matching (PSM) were conducted in three subject areas: Algebra I, English I, and Biology. PSM identifies a sample taking the paper test that is comparable to the sample taking the online test. As a result, it does not require that students take a test in both modes. The purpose of the follow-up studies was to explore the feasibility of using propensity-score matching as an alternative to requiring examinees to take two tests. For these three subject areas, results from comparability analyses from PSM were compared to results from analyses based on the more burdensome double-testing (within subjects) design. In Civics & Economics, propensity scoring was used as a complement to the within-subjects design because of the problems encountered in counterbalancing. A propensity score matching was not used in US History because of the results of that matching (as well as an exact match method) did not produce reasonable results when compared to the within-subjects results. The poor matching results suggest that the covariates used were not capturing key differences between the paper and online groups.

The PSM data were used to analyze most aspects of comparability. The relationship between mode-based scores (such as correlations or agreement in achievement levels) was not examined with PSM data because these values could not be computed with between-subjects data. Rather, within-subjects data were used. In addition, additional propensity score studies were conducted in Algebra I in an attempt to improve upon the method. Table 1 lists the types of comparability studies conducted for this grant and references to the reports written. More detail on each of the studies can be found in the reports.

Table 1. Types of Comparability Studies Conducted on the End of Course Data

Subject Area	Within-subjects	Propensity Score	Counter-Bal. successful?	Reports
Algebra I	Yes	Yes	Yes	Lottridge, Nicewander, and Mitzel (2008a) Lottridge, Nicewander, and Schulz (2008a) Lottridge, Nicewander, and Box (2008)
English I	Yes	Yes	Yes	Lottridge, Nicewander, and Mitzel (2008b) Lottridge, Nicewander, and Schulz (2008b)
Biology	Yes	Yes	No	Lottridge, Nicewander, and Mitzel (2008c) Lottridge and Nicewander (2008a)
Civics & Economics	Yes	Yes	No	Lottridge and Nicewander (2008b)
U.S History	Yes	No	No	Lottridge and Nicewander (2008c)

Examinees

Examinees were students who were enrolled in one of the designated courses and who were required to take the state's End-of-Course test in Spring 2007 in one of the five subject areas. The study sample was limited to students who: a) commonly took that course (e.g., 10th and 11th graders); b) had tests with no administration issues; c) attempted at least one item on the test; d) had an online test with a valid administration date; and e) did not attend an alternative school. The full sample for the within-subjects study consisted of all students with two valid scores, one on the paper test and one on the online test. The sample for the PSM (between-subjects) study consisted of students who took the online test first (the treatment group) and a matched group of

students, extracted from the population of students of who took only the paper test (the control group).

For each subject area, Table 2 displays the number of schools that participated in the study, the number of examinees overall, the number of examinees taking the paper test first, the number of examinees taking the online test first, and the number of examinees in the propensity score matching study. In Algebra I, three propensity score studies were conducted:

- Study 1 conducted the matching of the paper and online samples across grades 8 and 9, and did not impute missing data.
- Study 2 conducted matching separately for grades 8 and 9 and did not impute missing data.
- Study 3 conducted matching separate for grades and used mean imputation on missing data.

Study 2 produced results most similar to the within-subjects and these results were used throughout the remainder of this report.

Table 2. Sample Sizes and Number of Schools of Comparability Studies Conducted on the End of Course Data

Subject Area	Within-Subjects			Online First Examinees	Propensity Score Online First/Matched Paper Examinees ¹
	Total Schools	Total Examinees	Paper First Examinees		
Algebra I	49	2101	1202	899	Study 1: 741 *Study 2: 741 Study 3: 899
English I	25	1527	702	825	536
Biology	20	1004	194	810	764
Civics & Economics	19	951	272	679	579
U.S History	17	938	317	621	n.a.

Note. n.a. = not applicable. * = Study results used in this report. ¹ Students who took the online test first were matched to students who had only taken the paper-based test.

Procedures

The tests were administered in Spring 2007 as part of operational testing for the End-of-Course program. The state department of education enlisted volunteer schools to participate in the study. The department made a concerted effort to obtain a representative sample of the state. Schools' motivations for participating in the study were: 1) The opportunity for students to be tested twice with the highest test score counting; 2) Preparation for the eventual use of statewide online testing; 3) A sufficiently long testing window; and 4) entreaties by the department. Presumably, schools that chose not to participate were not comfortable with online testing or did not want the additional testing burden required by the within-subjects design. The state used random assignment at the school level to counterbalance test administration order. Schools with limited computer access were given the option of dividing their students into the two test administration orders.

A maximum of two weeks was allowed between test administrations. Students were told that the higher of the two test scores counted as the official test score in order to help foster similar motivation levels across administrations.

Test Instruments

Prior to conducting the comparability studies, the state had created multiple test forms to assess student achievement in each of the five subject areas. These forms were designed according to blueprint specifications aligned to state-specified goals for that course. All forms consisted of multiple-choice items.

The test forms were originally created for paper-based administration, and were later adapted for online administration. The online testing software presented items on the screen one at a time-- a few items, with accompanying stimulus material, required scrolling. The examinee had the option of increasing or decreasing the font size, navigating around the test at will, and clicking on an answer choice or typing in a letter. Forms were spiraled in the paper-based and online administrations.

Table 3 displays the number of items on each test, the number of paper forms, the number of online forms, the number of forms administered in both modes, and whether forms had overlapping items. Because forms were spiraled in both administrations, some examinees took the same form in both modes. Because some forms shared items, some examinees took forms that shared items. The extent of form overlap was 20, 40, and 60 items, depending upon the subject area.

Table 3. Number of Test Forms in Each Mode and Form Overlap Across Modes, for Each Subject Area

Subject Area	Test Length	Number of Paper Forms	Number of Online forms	Number of Forms Administered in Both Modes	
				Number of Forms Administered in Both Modes	Form overlap?
Algebra I	64	5	5	5	Yes
English I	56	6	6	6	No
Biology	88	3	1	1	No
Civics & Economics	80	5	5	4	Yes
U.S History	80	5	5	3	Yes

Counterbalancing Analyses

The analyses of test order suggested that some schools did not adhere to their assigned test order. Because random counterbalancing is critical to control for a test order effect, it was important to investigate whether the schools that did not adhere to their assigned order differed from those who did. The order in which schools tested was verified by both a data-based analysis and by a survey of district or school test coordinators. Schools with an unverified or ‘mixed’ test order were removed from the sample.

In Algebra I and English I, the previous year's End of Grade test scores in math and reading, respectively, were used to determine whether there were differences among the samples of schools that adhered to test orders, did not adhere to test scores, and schools not assigned a test order. Scaled score and standardized scaled score mean differences were analyzed across the two groups. If the differences varied significantly across the three groups, then it was assumed that the samples did not come from the same population. Based on these analyses, only schools that adhered to their assigned test order were retained for analyses in Algebra I. Schools that adhered to their assigned test order or were not assigned a test score were retained for analyses in English I. In Biology, Civics & Economics, and U.S. History, examinee participation rates in the test administration orders suggested that counterbalancing was not implemented as expected. For these subject areas, further counterbalancing analyses were not conducted.

Analysis of form status

Examinees in the within-subjects design took either the same form, forms that overlapped on some items, or a completely different test form. For each subject area, within-subjects (mode) and between subjects (test order, form overlap status) ANOVAs were conducted to determine if the mean End-of-Course scaled scores differed across these factors. These analyses were undertaken primarily to examine whether there was a practice effect due to the degree of form overlap. In all subject areas, there was either a three-way interaction (English I, Algebra I) between mode, test order, and form overlap status or a two-way interaction (Biology, Civics & Economics, U.S. History) between form overlap status and test order. The statistical significance indicated that form overlap was an important effect, and that larger mean differences between modes were associated with higher degrees of overlap. These results are consistent with a practice/memory effect. As a result, the analyses were disaggregated by the extent of form overlap.

Propensity score matching

Propensity score matching was investigated as an alternative method of evaluating score comparability of the online and paper tests. Such a procedure allows a state to use the data from the regular online and paper administrations rather than administering the two types of tests to the same sample of students. The results from the online administration are compared to results from a matched sample from the regular paper administration.

PSM (Rosenbaum & Rubin, 1983) attempts to predict group membership using logistic regression, with the covariates as predictors. A one-zero variable, indicating whether a person is in the treatment (online) or control (paper) group, is regressed on the covariates using logistic regression. The so-called propensity score for each person is the probability of being in the treatment group. Each person in the treatment group is matched to an accompanying person in the control group using a "nearest neighbor" (in terms of the propensity score) from the control group. Once the PSM procedure is complete, the two matched groups can be compared on one or more independent variables.

PSM represents an improvement over exact matching methods because it

- contains exact matching information in the propensity score (e.g., equal propensity scores suggest an exact match);
- allows for matching on a single value, the propensity score;
- uses statistical significance to identify characteristics that predict group membership; and,
- allows the statistical model to determine closest matches when exact matches are not possible.

As with any other matching method, the quality of the match is determined by the availability and quality of matching variables. Weaknesses specific to PSM are that there are several steps in the matching procedure, and it can be difficult to determine the quality of the match without comparing the matched groups on key variables.

Propensity score matching was used as a follow-up to the Algebra I, English I, and Biology within-subjects studies. It was used as a complement to the within-in subjects design in Civics & Economics because counterbalancing was not implemented as anticipated. In Algebra I, PSM was studied more in depth by separating the matching by grade and by imputing missing data. More details on PSM as applied to these studies can be found in Lottridge, Nicewander, and Schulz (2008a), Lottridge, Nicewander, and Schulz (2008b), Lottridge, Nicewander, and Box (2008), Lottridge and Nicewander (2008a), and Lottridge and Nicewander (2008b).

Sample characteristics

Characteristics of the within-subjects sample were compared with characteristics of the examinees who did not participate in the study and took only the paper test. Individual, school, and test administration characteristics were compared on a large number (14 to 17) of characteristics. Individual characteristics compared were sex, ethnicity, grade level, free lunch status, LEP status, exceptionality, and related test scores (if available). School characteristics compared were school type, Title I status, region of state, and wealth rank of the county. Test administration characteristics compared were testing cycle, amount of make-up testing, and accommodations. The purpose of these comparisons was to provide evidence that the sample results and conclusions can be generalized to the entire population.

The within-subjects samples were similar to the population on most variables. The differences are outlined in Table 4. The major differences across subject areas were that examinees in the study sample came from different regions of the state and generally poorer counties. Additionally, the online sample tended to include more charter school examinees than the paper sample.

Table 4. Characteristics (of 14-17 included in the comparison) in which the Sample Differed from the Population in the Five Subject Areas

Subject Area	Characteristics
Algebra I	<ul style="list-style-type: none"> • 8th graders, academically/intellectually gifted, and students enrolled in charter schools were over-represented in the sample • 9th graders, students receiving free lunch, students not identified as exceptional, students enrolled in ‘regular’ schools and in non-Title I schools were under-

Subject Area	Characteristics
English I	<ul style="list-style-type: none"> represented in the sample • Students in sample had a moderately higher mean computer skills score • Students in the sample came from proportionally different areas of the state and came from slightly poorer counties • Males, Blacks, students receiving free lunch, students enrolled in charter schools and in schools with school-wide Title I status were over-represented in the sample • Females, Whites, students paying full price for lunch, and students enrolled in ‘regular’ schools were under-represented in the sample. • Students in the sample had slightly lower mean reading, computer skills, and math scores • Students in the sample came from proportionally different areas of the state and came from poorer counties
Biology	<ul style="list-style-type: none"> • 10th graders, students receiving temporary free lunch, ‘non-exceptional’ students, and students enrolled in charter schools were over-represented in the sample • 11th graders, students paying full price for lunch, and students enrolled in ‘regular’ schools were under-represented in the sample • Students in the sample came from proportionally different areas of the state and came from poorer counties
Civics & Economics*	<ul style="list-style-type: none"> • 10th graders, students receiving free lunch, students not LEP, ‘non-exceptional’ students, students enrolled in charter schools and non-Title I schools were over-represented in the sample • Students paying full price for lunch and students enrolled in ‘regular’ schools were under-represented in the sample • Students in the sample came from proportionally different areas of the state and came from poorer counties
U.S History	<ul style="list-style-type: none"> • Females, 11th graders, whites, ‘non-exceptional’ students, students enrolled in charter schools and in non-Title I schools were over-represented in the sample • Students paying full price for lunch and students enrolled in ‘regular’ schools were under-represented in the sample • Students in the sample had slightly lower mean reading and math scores • Students in the sample came from proportionally different areas of the state and came from poorer counties

*=The sample taking the online test first was used for this set of comparisons.

Overall methodology

Tables 5 and 6 outline the methods used in the comparability study designs, as well as the impact of those methods on interpretation of results. Table 5 displays methods and impacts relative to external validity evaluations, and Table 6 displays this information relative to internal validity evaluations. Sampling, instrumentation, administration, and scoring were used as organizers in considering the impact of design on validity issues.

The major external validity issues related to the volunteer sample. Volunteer samples are the only feasible method for conducting studies in educational research, but self-selection in participation weakens the external validity argument. The comparison of the sample and general

population yielded some differences, particularly related to the region of the state and the wealth of the counties. The major internal validity issues related to the lack of successful counterbalancing in three subject areas, and the small sample size for each form which precluded IRT analyses. IRT analyses would have enabled the comparison of test characteristic curves and number correct to scaled score tables. These comparisons provide the most detailed evidence of score comparability. However, the comparability studies, as conducted, adequately addressed each of the issues outlined in Tables 5 and 6.

Table 5. Potential Effects on the External Validity of the Mode Comparability Studies

Category	Design	Potential impact on external validity
Sampling		
Identifying the population	Sample restricted to 'typical' students taking the course and test in Spring 2007.	Results may not generalize to 'non-typical' students.
Obtaining the sample	<p>Schools volunteered to participate in the study.</p> <ul style="list-style-type: none"> • Reasons to participate: <ul style="list-style-type: none"> ○ Higher test score used ○ Practice for future online testing ○ Entreaties by state department • Reasons not to participate: <ul style="list-style-type: none"> ○ Resources to test students twice ○ Lack of technology resources for online testing 	Because schools were not chosen randomly, schools that participated may differ from the population. For example, schools confident in their technology use may be more likely to participate than schools that are not confident.
Sampling unit	Sampling was at school level. Schools were randomly assigned test order.	Sampling at the school level may create non-equivalent groups because it can be difficult to obtain school-based samples that exactly match on key characteristics.
Sample participation	Most schools who were assigned a test order chose to participate. Some schools did not adhere to assigned test order. Reasons for non-participation were not gathered. Schools with unverified test order were removed from sample. In some cases, schools that did not adhere to their assigned test order were removed.	Self-selection in participation and to adherence to assigned test order may result in a sample unlike the population. For instance, schools may choose not to participate if they were assigned a test order that they felt disadvantaged their students. Or, schools may ignore their assigned test order and choose their own test order.
Similarity to population	The sample and population were compared on a large range of student, test, administration, and school characteristics.	Differences between the sample and population may reduce generalizability. Overall, the samples were reasonably close to the population. Primary differences were: distribution among regions of the state and mean county wealth rank.
Instrumentation		
Instruments used	Instruments studied were operational test forms.	None.
Administration		
Administration conditions	Tests were administered as part of operational testing. The higher of two test scores counted for final score. A maximum of two weeks were allowed between paper and online administrations.	Impact of taking two tests is unknown. Counterbalancing should average out the order effect.
Online experience	Students in the study did not receive training to take the online test. However, many students had the opportunity to take a computer skills test in 8 th grade.	It is possible that the state may offer training in the future, and this may affect the generalizability of the study results.

	Category	Design	Potential impact on external validity
Scoring			
	Scores used	The analyses examined raw scores, scaled scores, achievement levels, and proficiency categories.	None.
	Disaggregation	Mean scaled scores were disaggregated by subgroups such as sex and ethnicity to determine whether differential mode effects existed.	If differential mode effects were identified for subgroups, this may affect generalizability to these subgroups.

Table 6. Potential Effects on the Internal Validity of the Mode Comparability Studies

Category	Design	Potential impact on internal validity
Sampling		
Group assignment	Some schools did not adhere to their assigned test order. For some subject areas, the schools who were assigned to administer the paper first participated in lower numbers than schools who were assigned to administer the online test first.	For subjects in which counterbalancing was not successful in practice, the results were disaggregated by test order. This confounded the order and mode effects.
Sample size	The samples were edited to remove unverified schools. In some subject areas, the analyses were disaggregated by test order, form, and degree of form overlap (see Instrumentation).	The removal of examinees and disaggregation reduced the sample size sufficiently that some analyses (such as IRT-based analyses) could not be conducted.
Matching	Propensity score matching was used to identify a matched sample in the population to the online sample. The technique is new in comparability studies, and its feasibility is yet unknown.	Propensity score matching can only produce matches on covariates entered into the model. If these covariates do not adequately represent the sample and population, then the two samples may differ in key way, influencing the study results. The results of these studies suggest that PSM produces comparable samples.
Instrumentation		
Forms used	Forms were spiraled within the online and paper conditions, and as a result some examinees took the same form in both modes. In addition, some forms shared items, and as a result some examinees to a subset of the same items in both modes.	The practice/memory effect for students taking the same form or forms with overlapping items is likely to be stronger than that for students taking different forms across modes. This effect can be controlled by disaggregating results across degrees of form overlap.
Online experience	The online forms were paper forms adapted into the online environment. The online environment was developed to work similarly to the paper environment.	None.
Administration		
Administration conditions	Schools were allowed a maximum of two weeks to administer the tests in both modes. Data were not collected to determine the time elapsed between the two administrations for each school and student.	Differences in the time elapsed between the test administrations may influence the extent of the order effect. For these studies, it was assumed to be random error.
Study conditions	No major issues were reported with the online testing.	Issues related to computer-based testing (e.g., crashes) can be considered random error.
Scoring		

	Category	Design	Potential impact on internal validity
	Scores used	The paper-based number correct to scaled score tables were used to assign online responses a scaled score. This method assumes that the IRT item parameters and TCC of the online test are (within measurement error) the same as that of the paper test.	The use of the paper-based number correct to scaled score tables may obscure mode effects. However, if the mean and standard deviation of the scaled scores are the same across the paper and online groups, then it may be assumed that no mode effect has occurred at the scaled score level. If the mean and standard deviation of the scaled scores differ both statistically and practically, then a mode effect has occurred.
	Score comparisons	When no paper data were available, the online results were compared to expected (from IRT parameters) paper results in order to better understand the mode effect.	This approach helped to better understand the direction and magnitude (if any) of the mode effect.

Analyses

Chapter 1 describes four criteria that can be used to judge whether test variations are comparable: the extent to which the online test produces more valid inferences, addresses the same content standards, is equally as reliable as the paper test, and classifies examinees in the same manner. The first criterion (producing more valid inferences) is not relevant to the studies presented in this report; the online tests were simply computer-based versions of the paper tests. One could imagine that this criterion is relevant for other types of online tests. In particular, online tests that use innovative items in order to more accurately assess content standards may be examined against this criterion.

Using the figure from Chapter 1 regarding the content and score dimensions of comparability, the comparability examined in these studies represents the highest expectation of comparability. Because the online forms in this series of studies are online versions of paper forms, we expect to be able to judge comparability at the item level and at the raw score level.

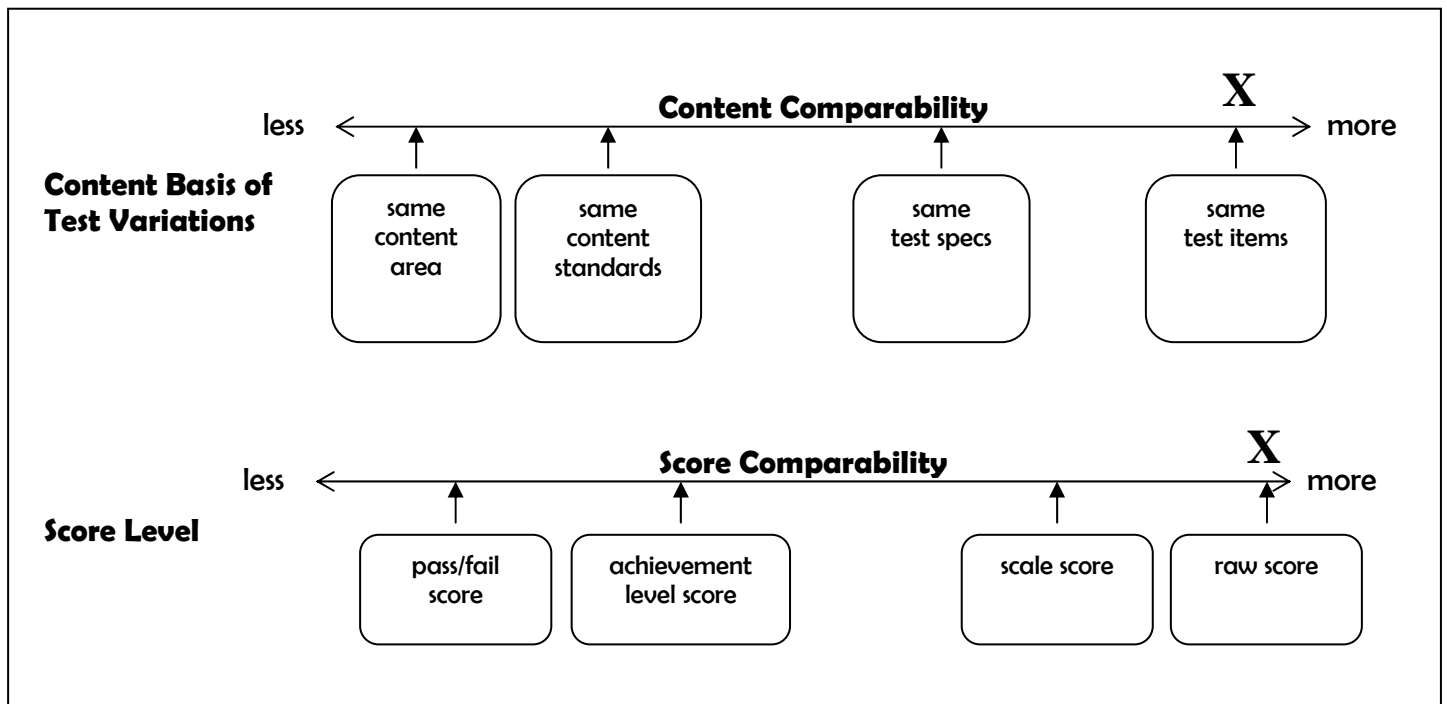


Figure 1. Comparability continuum

Table 7 presents the three criteria, suggested analyses, and the analyses conducted in the five studies. Not all of the suggested analyses were conducted in these studies. Because the sample sizes across conditions (*i.e.*, across forms) were small, IRT analyses could not be conducted for four of the five subject areas. This meant that conditional standard errors of measurement, test characteristic curves, number correct to scaled score tables, and item-level dimensionality tests could not be conducted. A large enough sample was available in the Biology PSM study to

conduct some IRT analysis; the test characteristic curves and mean item parameter estimates were compared across the modes.

Table 7. Study Design and Analyses by Three Comparability Criteria

Comparability Criterion	Suggested Analyses	Analyses Conducted
Assess Same Content Standards	Test blueprint comparisons	Yes
	Expert review of item online modifications	No
	Comparisons of different test formats	No
	Tests of dimensionality at item level	No
	Tests of dimensionality at test level	Algebra I, English I
	Item level comparisons / DIF	Yes
Comparable Reliability	Overall reliability values	Yes
	Conditional SEM / Test information	No
Comparable Classification	Frequency distribution of scores	Yes
	Measures of central tendency and dispersion	Yes
	Cumulative frequency distribution	No
	Test characteristic curves	Biology
	Number correct to scaled score tables	No
	Performance level distributions	Yes
	Correlation (corrected and uncorrected for unreliability)	Yes
	Agreement of assignment into performance levels	Yes
	Correlation (corrected and uncorrected for reliability) with variables	Yes
	ANOVA by subgroups (e.g., gender and race)	Yes

Results

The three relevant comparability criteria were used to organize the study results. In general, the results across the five studies suggest that the online and paper tests appear to be measuring similar content standards with the same level of reliability, and are slightly more difficult than their paper counterparts. The difference in difficulty suggests that there is some construct irrelevance variance associated with the online version; however, the effect of that variance is small and does not appear in analyses aside from mean score differences. In essence, the results suggest that once equating is conducted to remove the difference in difficulty, the online and paper versions of the tests are comparable. The follow-up propensity score matching results mirrored the within-subjects results in Algebra I, English I, and Biology, suggesting that PSM is a feasible method for identifying matched samples in comparability studies. A brief overview of the study results appears below for the three criteria.

The online and paper versions should address the same content standards.

The online test forms were paper forms adapted to the online environment. As a result, the online forms and paper forms had the same test blueprint specifications and were comprised of the same item text. The online environment was created to be similar to the paper environment, as described in the methods section.

The overall test factor structure was examined only for Algebra I and English I. These subjects were chosen because data from a previous year's subject test could be used to satisfy the model constraints. Additionally, the factor structure could only be compared in a within-subjects study. A confirmatory factor analysis (CFA) was conducted with increasing levels of constraints to test whether the paper and online test were parallel, tau-equivalent, or congeneric. The results of the CFA supported the parallel test model.

Differential item functioning (DIF) analyses were also conducted in all studies to determine whether, controlling for ability, examinees were likely to perform better or worse on test items in one or the other mode of administration. The Mantel-Haenszel test was used, along with a classification of effect size used by the Educational Testing Service (ETS; Zwick and Ercikan, 1989). Very few items (less than 4.4 % across all subject areas) were identified as exhibiting DIF. Follow-up PSM DIF studies produced the same result, although different DIF items were identified between the within-subjects and follow-up DIF study.

The online and paper scores should be equally reliable

The internal coefficient of reliability (Cronbach's alpha) was calculated based on raw score data for each online and paper forms in each subject area. In the within subjects analyses for Algebra I, English I, Biology, and US History and the PSM analysis conducted for Civics & Economics, the reliability values were very close, within .03 of one another. In the follow-up PSM analyses for Algebra I, English I, and Biology, the reliability values were also very close, again within .02 of one another.

The online and paper versions should classify students similarly.

The classification criterion covers a range of statistical analyses, including mean comparisons, cross-mode correlations, decision consistency analyses, and sub-group analyses. The results are presented for each of these analyses.

Mean differences

Table 8 presents the mean scores in each mode for the five subject areas. In Algebra I and English I, the paper first and online first groups were combined because counterbalancing was implemented as intended. In Biology, Civics & Economics, and U.S. History, counterbalancing did not work as intended. For Biology and U.S. History, the data are presented separately for the online first and paper first students. Only propensity score matching was conducted for Civics & Economics. Only data from students taking different forms of the test across modes are included in the within-subjects analysis. Follow-up propensity score matching results are presented for Algebra I, English I, and Biology.

Generally, mean differences across all forms in each subject area were small to moderate -- less than 2 scale score points. The standardized mean differences ranged from -.25 to +.16. In Algebra I, English I, Biology, and Civics & Economics, the online test was slightly harder than the paper test. The follow-up PSM mean differences were similar to the within-subjects results for Algebra I and English I. The follow-up PSM differences in Biology were larger than were seen in the within-subjects mean differences, and suggested that the online test was harder than the paper test. In U.S. History, the mean online test scores were 1.4 points higher than the paper test scores when the paper test was administered first, and almost the same when the online test was administered first. Both the Biology and U.S. History within-subjects results are consistent with a practice effect; the score of the second test was higher than the score of the test taken first.

Table 8. Scaled Score Mean and Standard Deviations and Differences by Subject Area and Study Design

Subject Area	Study Design	Paper Mean (SD)	Online mean (SD)	Mean difference (online – paper)	Standardized mean difference
Algebra I	WS:	154.61 (10.37)	153.34 (10.53)	-1.27	-0.12
	PSM:	155.05 (10.84)	153.99 (10.97)	-1.06	-0.10
English I	WS:	149.44 (8.46)	147.72 (8.40)	-1.72	-0.20
	PSM:	150.59 (8.64)	148.44 (8.09)	-2.15	-0.25
Biology	WS PF:	57.44 (6.53)	57.89 (6.31)	+0.45	+0.07
	WS OF:	56.47 (6.73)	55.95 (6.52)	-0.52	-0.08
	PSM:	57.34 (7.04)	55.77 (6.67)	-1.57	-0.22
Civics & Economics	PSM:	151.91 (8.98)	150.82 (8.09)	-1.06	-0.12
U.S History	WS PF:	147.56 (8.80)	148.96 (9.19)	+1.40	+0.16
	WS OF:	149.92 (9.31)	149.85 (8.93)	-0.07	<-0.01

Note. WS=Within-subjects; PSM=Propensity score matching; PF=Paper first administration; OF=Online first administration.

Cross-mode correlations

The correlations between paper and online scaled scores were computed in the five studies, as were the agreements in achievement level assignment. The correlations of overall paper and online scaled scores ranged between .80 to .90 for all subject areas. The corrected correlations ranged from .87 to .97. Table 9 presents the correlation results for the five studies. Correlations were corrected by the average alpha reliability across forms in each mode.

Table 9. Corrected and Uncorrected Correlations of Paper and Online Scaled Scores

Subject Area	Design	Correlation	Corrected Correlation
Algebra I	Test orders combined	.90	.97
English I	Test orders combined	.82	.91
Biology	Paper first	.84	.89
	Online first	.84	.89
Civics & Economics	Paper first	.80	.87
	Online first	.88	.96
U.S History	Paper first	.89	.97
	Online first	.87	.94

Decision consistency

Decision consistency between the online and paper-assigned achievement levels was also evaluated and compared to expected decision consistency rates of two paper tests administrations. Table 10 presents the observed exact agreement of achievement levels assigned using the online and paper scaled scores. In Algebra I, the observed exact agreement rate was almost identical to the expected agreement rates. In English, the observed exact agreement rate was 9.4% lower than the expected rate. This difference was presumably due to mean differences between the online and paper scores. The Biology observed exact agreement rates for the paper and online first samples were both similar to the expected agreement rate, suggesting that the mean differences were small. The Civics & Economics and U.S. History rates were much lower than the expected rates, and these differences were likely due to both a combined mode-practice effect.

Table 10. Observed and Expected Exact Agreement of Achievement Level Assignment

Subject Area	Design	Observed	Expected
Algebra I	Test orders combined	70.1%	70.8%
English I	Test orders combined	65.4%	74.0%
Biology	Paper first	71.5%	74.3%
	Online first	71.4%	
Civics & Economics	Paper first	65.2%	75.0%
	Online first	68.4%	
U.S History	Paper first	66.5%	73.5%
	Online first	67.2%	

Table 10 presents the exact agreement of proficiency level classification (i.e., Below Proficient vs. Proficient and Above) using the online and paper scaled scores. In Algebra I, the observed exact agreement rate was almost identical to the expected agreement rates. In English, the observed exact agreement rate was 5.5% lower than the expected rate. Again, this difference was presumably due to the mean differences between the online and paper scores. Unlike the

achievement level decision consistency rates, the Biology observed exact agreement rates were lower than the expected rates. The Civics & Economics and U.S. History rates were much lower than the expected rates, and these differences were likely due to a combined mode-practice effect.

Table 11. Observed and Expected Exact Agreement of Proficiency Level Classification

Subject Area	Design	Observed	Expected
Algebra I	Test orders combined	88.3%	87.7%
English I	Test orders combined	84.5%	90.0%
Biology	Paper first	82.1%	90.2%
	Online first	84.1%	
Civics & Economics	Paper first	82.6%	89.7%
	Online first	84.7%	
U.S History	Paper first	84.9%	88.9%
	Online first	86.8%	

Relationship to external measures

Correlations with various external measures were computed and compared for the paper and online scaled scores. If the online and paper versions are testing the same content standards, then their correlations with other measures should be similar. In Algebra I and English I, the external measures used were a computer skills test administered in 8th grade, an 8th grade math test, and an 8th grade reading test. In Biology and U.S. History, the external measures used were a 10th grade Algebra test, a 10th grade English test, and the student’s anticipated grade in the course. In Civics & Economics, an 8th grade reading test and the student’s anticipated grade in the course were used. Table 12 presents the correlations. None of the differences in correlations across mode were significant at the .05 level.

Table 12. Similarity of Paper (Online) Correlations with External Measures

Subject Area	Study design	Computer skills test	Math test	Reading/ English test	Anticipated course grade
Algebra I	Test orders combined	.68 (.68)	.84 (.84)	.69 (.70)	
	PSM	.72 (.69)	.86 (.86)	.72 (.70)	
English I	Test orders combined	.59 (.61)	.71 (.71)	.77 (.79)	
	PSM	.66 (.64)	.76 (.74)	.83 (.79)	
Biology	Paper first		.55 (.58)	.69 (.69)	.61 (.64)
	Online first		.61 (.59)	.73 (.67)	.57 (.54)
	PSM		.65 (.60)	.67 (.68)	.48 (.52)
Civics & Economics	PSM			.74 (.71)	.57 (.56)
U.S History	Paper first		.55 (.56)	.66 (.60)	.49 (.45)
	Online first		.48 (.48)	.59 (.66)	.54 (.54)

Note. Number in parenthesis is online correlation. * $p < .05$.

Subgroup differences

Analyses of variance were conducted in each study to determine whether there was a differential mode effect for members of subgroups. The subgroups examined were sex, ethnicity, grade, free lunch status, LEP status, and exceptionality status. Analyses were conducted only for subgroups

with two or more levels with more than 30 examinees. Within-subjects ANOVAs were conducted in the within-subjects studies and between-subjects ANOVAs were conducted in the propensity score studies. Within-subjects ANOVAs are more sensitive to differences, and so are likely to produce statistically significant results. The between-subjects ANOVAs were conducted on a smaller sample (online first examinees) and so significance was less likely to be detected and fewer groups were examined due to samples sizes being below 30 in subgroup levels.

Table 13 displays the results of the ANOVAs for each subject area, study design, and subgroup. ‘Yes’ in a cell means that statistical significance was detected for that subgroup and indicates a possible differential mode effect. ‘No’ means that statistical significance was not detected, and indicates no mode effect. Very few differential mode effects were detected overall. A differential effect for free lunch status was detected in Algebra I and Civics & Economics. A gender effect was detected for Biology, and a differential effect for ethnicity was detected for U.S. History. The propensity score results were mostly similar for Algebra I, English I, and Biology. In Algebra, the within subjects ANOVA detected a differential mode effect for free lunch status, but the PSM ANOVA did not. In Biology, within subjects ANOVA detected a differential mode effect for free lunch status, but the PSM ANOVA did not.

Table 13. Existence of Mode Effects by Subgroup

Subject Area	Study Design	Sex	Ethnicity	Grade	Free Lunch Status	LEP	Exceptionality
Algebra I	WS	No	No	No	Yes	No	No
	PSM	No	No	No	No	No	No
English I	WS	No	No	No	No	No	No
	PSM	No	No	n.a.	No	n.a.	No
Biology	CBT first	Yes	No	No	No	n.a.	No
	PSM	No	No	No	No	n.a.	No
Civics & Economics	PSM	No	No	No	Yes	n.a.	No
U.S History	CBT first	No	Yes	n.a.	No	n.a.	No

Note. Yes=Statistical significance detected at the .05 level. No=Statistical significance not detected at the .05 level. n.a.=Analyses not conducted due to small sample size.

Limitations of the Studies

In general, there were four limitations to the comparability studies. First, the study used volunteer samples. While the use of volunteer samples is unavoidable in applied research, it is unclear whether schools that participated were different in some important way than schools that did not. A comparison of characteristics at the student, test, and school level suggested that the study samples came from different regions of the state and from poorer counties. In addition, charter schools had a slightly higher representation in the sample than in the population. Second, the within-subjects study changes the nature of the test administration process because examinees take the test twice. The effect of taking the tests twice on the scores is unknown. Counterbalancing deals with the order effect by placing the order effect as an error occurring equally for both test orders. In the case of Biology, Civics & Economics, and U.S. History, counterbalancing did not work effectively, and so these analyses were disaggregated by test

order. Disaggregating the test results confounds the mode and order effect. Third, while the overall sample size was large, the sample size for each form was relatively small. The sample size was further reduced by schools not adhering to the assigned test order and by the fact that some examinees were assigned the same form. The small sample size meant that some analyses could not be conducted, such as analyses involving item response theory (e.g, calibration of online item parameters, computation of conditional standard errors), equating, or item-level factor analyses. Fourth, the propensity score matching procedure did not produce completely matched paper and online samples. Differences may have been due to the matching variables employed. In addition, the propensity score matching produced unreasonable results in U.S. History. A follow-up analysis showed that exact matching also produced unreasonable results, suggesting that the covariates used in the matching were not adequate for the matching procedure. The covariate set in U.S. History included the 10th grade English score as the achievement variable, and this variable had a fairly low correlation with the U.S. History scaled score (.59). It is possible that this achievement variable was not a sufficient proxy for U.S. History knowledge and skills. All other subject areas used an achievement variable with at least a correlation of .70 with the subject-area scaled score.

Conclusions

Overall, these analyses suggest that the paper and online scores are comparable using the three relevant comparability criterion. Scores in both modes appear to be measuring the same content standards, with the same level of reliability, and appear to be classifying examinees in mostly the same way. The mean online test scores were lower than the paper test scores, and more online examinees were placed in lower achievement categories. However, this mode effect appears to be pervasive across all test items, and does not appear to be influencing any particular subgroup. The difference in mean scores suggests that equating might be required to ensure that the scores are truly interchangeable.

The analyses also indicate that propensity score matching produced similar results and interpretations as the within-subjects studies for Algebra I, English I, and Biology. These results suggest that propensity score matching is a viable procedure for identifying matched samples to be used in comparability studies when a proper covariate set (particularly, a strong achievement variable) is available.

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2004). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 283-302.
- Lottridge, S.M. & Nicewander, W.A. (2008a). *Comparing computer-based and paper-based test scores in one state's End-of-Course biology program: Results using propensity score matching*. Monterey, CA: Pacific Metrics Corporation.

- Lottridge, S.M., & Nicewander, W.A. (2008b). *Comparing computer-based and paper-based test scores in one state's End-of-Course civics & economics program*. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., & Nicewander, W.A. (2008c). *Comparing computer-based and paper-based test scores in one state's End-of-Course U.S. history program*. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., Nicewander, W.A., & Box, C. (2008). *Comparing computer-based and paper-based test scores in one state's End-of-Course algebra I program: Results using propensity score matching with two approaches for missing data*. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., Nicewander, W.A., & Mitzel, H.C. (2008a). *Comparing computer-based and paper-based test scores in one state's End-of-Course algebra I program*. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., Nicewander, W.A., & Mitzel, H.C. (2008b). *Comparing computer-based and paper-based test scores in one state's End-of-Course English I program*. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., Nicewander, W.A., & Mitzel, H.C. (2008c). *Comparing computer-based and paper-based test scores in one state's End-of-Course biology program*. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., Nicewander, W.A., & Schulz, E.M. (2008a). *Comparing computer-based and paper-based test scores in one state's End-of-Course algebra I program: Results using propensity score matching*. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., Nicewander, W.A., & Schulz, E.M. (2008b). *Comparing computer-based and paper-based test scores in one state's End-of-Course English I program: Results using propensity score matching*. Monterey, CA: Pacific Metrics Corporation.
- Lottridge, S.M., Nicewander, W.A., Schulz, E.M., & Mitzel, H.C. (2008). *Comparability of paper-based and computer-based Tests: A review of the methodology*. Monterey, CA: Pacific Metrics Corporation.
- Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- Winter, P.C. (2009, this volume). XXXXXXXX [Chapter 1]. In Winter, P.C. (Ed.), xxxxxx [title of handbook]. Washington, DC: Council of Chief State School Officers.
- Zwick, R & Ercikan, K. (1989). Analysis of Differential Item Functioning in the NAEP History Assessment. *Journal of Educational Measurement*, 26(1),55-66