

Comparability of Paper-based and Computer-based Tests: A Review of the Methodology

Susan Lottridge, Ph.D.
Alan Nicewander, Ph.D.
Matt Schulz, Ph.D.
Howard Mitzel, Ph.D.

**Submitted to the
CCSSO Technical Issues in Large Scale Assessment Comparability Research Group**

January 3, 2008



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

DRAFT

DRAFT

DRAFT

Comparability of Paper-based and Computer-based Tests: A Review of the Methodology

Introduction

The number of computer-delivered tests is increasing in schools as a result of accountability legislation and the perceived advantages of computer-delivered tests. In 2003, an *Education Week* survey reported that thirteen states were piloting or using computer-based tests (Olson, 2003). In 2006, twenty-one states and the District of Columbia offered computerized tests in some form (Educational Research Center, 2006). Perceived advantages of computer-based testing include more flexible scheduling, the ability to tailor tests more specifically to student needs, and more rapid scoring and reporting.

Comparability issues arise when a new mode of test delivery replaces or is used alongside an established mode, such as paper-and-pencil testing. Because paper-and-pencil tests (PBTs for paper-based tests or testing) have precedence of use, they represent the gold standard to which computer-based tests or testing (CBTs) is compared. As PBTs are replaced in the K-12 testing arena by CBTs, the CBT scores need to be tied back in some way to the original PBT scale in order to maintain continuity. Also, the development of a CBT-only test for which there is no PBT precedent within a given state or school district may nevertheless require consideration of its relationship to a paper-based counterpart in another district or state.

This report is intended to be useful to educational policy makers and researchers concerned with the comparability of computer-delivered and paper-and-pencil tests. The purpose of this report is to review the *methods* used in the literature to investigate comparability rather than the results of such studies. Excellent reviews of comparability study results are already available in the literature (Paek, 2005; Bennett, 2003; Gaskill, 2006).

Comparability

Comparability can be examined on two levels. First, comparability can be examined in terms of score equivalence. In other words, one can investigate whether the two modes (PBT or CBT) produce similar score distributions, such as similar means and standard deviations. Second, comparability can be examined in terms of construct equivalence. Here, the term “construct” refers to an unobservable property of persons that is being measured using a test. An example of a construct is math proficiency; a person’s proficiency at math cannot be directly observed but a person can be thought to have a level of this skill. Constructs can be narrowly defined (e.g., keyboarding proficiency) or broadly defined (e.g., reading comprehension). Because construct comparability involves determining whether the tests in two modes are measuring the same construct to the same degree, it is a complex and difficult task.

Why examine score equivalence? If the distribution of the scores is the same across the two modes, then two important inferences logically follow. First, the two modes can be said to be functionally comparable in terms of overall scoring of a sample. Second, the constructs measured can be *reasonably* assumed to be the same; there is no counter evidence that differing constructs

are involved when the score distributions are comparable. In fact, meta-analyses on comparability generally indicate that mode differences, if they exist, are small for untimed tests (Bergstrom, 1992; Mead & Drasgow, 1993; Kim, 1999). However, comparing score distributions alone may be misleading. Tests from two modes might produce the same overall score distributions, but the scores for any one examinee may differ substantially between tests. In other words, the scores from two modes might produce a different rank ordering of examinees. Such an occurrence is an indication that the modes are measuring different constructs. Using methods to examine the validity of the scores, one can compare the extent to which scores are measuring the same construct.

Score equivalence and construct equivalence are especially important for accountability. It would be difficult or impossible to monitor trends in student learning for purposes of accountability if scores from two alternative modes of assessment had different meaning, or if students in a particular demographic category tended to earn different scores on one mode than on the other. The meaning of change in a summary statistic, such as the percent of students scoring at or above a given achievement level, could be confounded by whether or not the test was administered in a paper-based or computer-based mode.

Various guidelines have been published for examining comparability between CBT and PBT. Standards from the American Psychological Association (1986) and the International Test Commission (2005) emphasize the need for similar score distributions, reliabilities, ranking of examinees, and correlations with external criteria.

“Scores from conventional and computer administrations may be considered equivalent when (a) rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode” (APA, 1986, p. 18).

“Provide clear documented evidence of the equivalence between the CBT/Internet test and non-computer versions (if the CBT/Internet version is a parallel form). Specifically, to show that the two versions: have comparable reliabilities; correlate with each other at the expected level from the reliabilities; correlate comparably with other tests and external criteria; and, produce comparable means and standard deviations or have been appropriately calibrated to render comparable scores” (ITC, 2005, p. 21).

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2004) provide less specific guidance on statistics to be used, and instead instruct the test developer to conduct studies relevant to the use and interpretation of the test scores.

Standard 4.10: “A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably. In some cases, direct evidence of score equivalence may be provided. In other cases, evidence may come from a demonstration that the theoretical assumptions underlying the procedures for establishing score comparability have been sufficiently satisfied. The specific rationale and

the evidence required will depend in part on the intended uses for which score equivalence is claimed” (p. 57).

Together, these guidelines call for comparisons of score distributions across testing modalities, comparisons of relationships of scores across modes, and also comparisons of relationships with other criterion measures. While the guidelines outline methods for investigating comparability, they do not specify criteria by which to determine whether comparability has been achieved. Rather, the investigator must use judgment when interpreting results. The remainder of this report will describe methodologies for making such comparisons, and will outline the work other researchers have conducted to determine comparability.

Use of the Principles of Construct Validation in the Design of Comparability Studies

The area of psychology from which one may draw methods for addressing comparability is *construct validation* (Cronbach & Meehl, 1955). Construct validation is the process by which evidence is gathered to assess the degree to which a test is measuring the construct it is intended to measure and the degree to which test scores are supporting intended inferences. The validation procedure consists of the accumulation of positive evidence *and* the inability to uncover negative evidence. At some point, enough evidence is acquired that enables a judgment regarding the validity of the test scores.

In terms of the philosophy of science, construct validation falls under the general topic of *theory testing*. Construct validation uses theory and/or logic to develop hypotheses that should be true if a test is valid for the construct that it claims to measure. These hypotheses form predictions that are tested using experimental data. In the case of comparability studies, the construct validation paradigm is simplified a bit since the nature of the construct being measured by two tests (or two testing modes) does not have to be identified. Rather, the question is whether the constructs assessed by the two tests are the same. In the determination of the degree of consonance between the constructs measured by CBT and PBT, the following is a partial list of the logical deductions that can be derived and then tested with experimental data:

- If the construct measured by the two modes is the same, then their content and content specifications should be the same (evidence from content validation).
- If identical constructs underlie these two testing modes, then they should have the same factor structure (psychometric-statistical evidence).
- If CBT and PBT measure the same construct—and are to be used interchangeably—they should have the same measurement precision.
- If the constructs being measured are the same, then these two testing modes should yield scores that differ only because of difficulty. This difference can be then removed using equating (psychometric-statistical evidence).
- If the underlying constructs are identical for two testing modes, then their intercorrelation, corrected for unreliability, should be unity—within sampling error.

This evidence would establish that the tests are *congeneric* (i.e., have perfectly correlated true scores); it does not confirm that they are parallel forms (i.e., yield identical true scores and the same error variance) (psychometric-statistical evidence).

- If the CBT and PBT measure the same construct, then they should have the same predictive validity coefficient (evidence from predictive validation), or, similarly, tests measuring the same construct should have equal correlations with external measures (concurrent validity).

Of course, the support (or lack thereof) for these hypotheses involves human judgment, probabilistic reasoning, and the strengths and limitations of study design. Thus, the interpretation of evidence is crucial to a decision regarding comparability. The key idea here is that comparability can and should be addressed experimentally using a hypothesis-testing approach in a construct validation framework.

Comparability Research Methodology Review

The purpose of this report is to provide an overview of the methods used in studies investigating comparability between PBT and CBT. The list of studies examined appears in Table 1. The focus of the comparability studies centered on score and/or construct differences between PBTs and CBTs, and all studies conformed to one or more of the six construct validation hypotheses listed above. No studies examined all six hypotheses, but many examined at least two. The most comprehensive studies considered the key aspects of validity and reliability, taking into consideration both score distributions and the relationships of CBT and PBT scores to the construct being measured.

Table 1. List of Studies Included in Methodological Review

Study
Choi and Tinkler (2002)
Eignor (1993)
Fitzpatrick and Triscari (2005)
Higgins, Russell, and Hoffman (2005)
Hollenbeck, Tindal, Stieber, and Harniss (1999)
Johnson and Green (2006)
Olson, Maynes, Slawson, and Ho (1989)
Poggio, Glasnapp, Yang, and Poggio (2005)
Pommerich (2004)
Pomplun and Custer (2005)
Pomplun, Frey, Becker, and Hughes (2000)
Russell and Haney (1997)
Russell and Plati (2001)
Russell and Tao (2004)
Russell (1999)
Sandine, Horkay, Bennett, Allen, Braswell, Kaplan, and Oranje (2005)
Schwarz, Rich, and Podrabsky, (2003)

Study

Way and Fitzpatrick (2006); Way, Davis, and Fitzpatrick (2006); Keng, McClarty, and Davis (2006)

Zhang and Lau (2006)

This section outlines the various approaches used in the literature that can serve as examples of potential comparability designs. Each design description will be accompanied by potential threats to the validity of the results. Only studies using K-12 students, having sufficient detail to understand the methodology employed, and incorporating reasonably strong design and implementation are presented.

General Methodological Issues

Study design issues generally span four major dimensions: sampling, instrumentation, administration, and scoring. The comparability literature (including but not limited to studies listed in Table 1) reviewed often did not provide sufficient detail or exhibited problems related to one or more of these dimensions. Most design issues arose due to real-world constraints in implementation; however, some were also due to inadequate descriptions of the implementation. The dimensions listed below are organized by the internal and external validity framework provided in the work of Campbell and Stanley (1963). External validity refers to the extent to which the study results can be generalized to the population. Internal validity refers to the extent to which the study design sufficiently controls variables to test the hypotheses.

General factors influencing external validity (generalizability of results). Each of the four dimensions of sampling, instrumentation, administration, and scoring are presented as they relate to the generalizability of the study results.

- Sampling. The selection and solicitation of the sample should be described. This description should describe how the original sample (e.g., schools and/or students) was selected, how many of the sample participated in the study, and the extent to which the study participants were similar to the original sample. In particular, with voluntary samples it is important that attention be paid to the type of schools that volunteer since participating schools may differ from non-participating schools. For instance, a volunteer school may have more access to computers or may consist of more computer-savvy students.
- Instrumentation. The instruments used in the study should be comparable to those which would be used operationally. In the case of K-12 testing, this issue is less of a concern since it is likely that operational forms would be used in any comparability study. Studies which use researcher-assembled instruments need to provide details on test specifications, test construction, reliability, and, if possible, validity.
- Administration. The conditions under which the tests are administered should be similar to those normally practiced. These conditions include having the same motivation level (e.g., low-stakes versus high-stakes testing), similar computers and computer access, similar bandwidth issues, similar student practice test and training, and similar proctor training.

- Scoring. The method of scoring should reflect the methods used operationally. For instance, if examinees in the population are assigned into performance levels based on test scores, then the sample examinees should be similarly assigned.

General factors influencing internal validity (interpretation of results). Issues relating to the four dimensions and whether their influence on test scores could be attributed to mode and not other influences, are listed below.

- Sampling. The key areas of sampling with respect to internal validity are sample size and assignment to groups.
 - Sample size. It is important to collect a large enough sample in order to reflect the breadth of the population, compute stable statistics, and use the sample for future analyses (such as equating studies or IRT item calibrations). For instance, calibration of IRT item parameters is often an element of comparability studies. For the three-parameter IRT model, a minimum of 1,200 examinees is recommended for accurate calibration. For the Rasch model, a minimum of 500 observations is recommended. Equipercentile equating can require up to 5,000 examinees.
 - Group assignment. A key element of an experimental study is random assignment to help ensure group equivalence. Random assignment can occur at the group level (school, classroom) or student level. The equivalence of the groups should be demonstrated using background variables, such as demographic or testing data. Group assignment into counterbalanced conditions should also be random and not left up to the study participants. If random assignment is not possible, then efforts should be made to ensure the groups are similar (e.g., matching methods, use of covariates).
- Instrumentation. Instruments need a sufficient number of items to represent the construct of interest and to be reliable measures. This issue is particularly salient when using constructed-response items, because generally a small number of items are used and examinees often respond differently to similar-looking items. For instance, a student may perform and score differently on two items designed to be equivalent. As an example, two story problems might use the same underlying math but apply it to different scenarios. A student might know the underlying math, but be able to apply the computation to only one scenario. This examinee-by-task interaction is important to consider if different items are administered across mode since the interaction, rather than mode, may be the source of score differences.
- Administration. Tests of the two populations should occur at the same time whenever possible to rule out extraneous variables (e.g., maturation). Studies that use a volunteer sample and compare the sample results to already existing test data risk extraneous factors influencing test scores.
- Scoring. The test scores need to be placed on the same scale prior to any comparison. This linking process should be reported. In addition, if IRT item parameters are used in CBT or computer adaptive tests (CAT), then the origin of those parameters should be described. For instance, the item parameters may be from scores gathered from a PBT or items may be recalibrated on the CBT examinees. In addition, inter-rater reliability of constructed-response measures should be examined and presented. It is also important to identify which type of score is being used in any comparison (e.g., raw score, standard score, IRT-estimated theta, IRT-estimated true score).

Types of designs

Two basic designs are dominant in the comparability literature: the within-subjects design and the between-subjects design. In the within-subjects design, examinees take both a CBT and a PBT. In the between-subjects design, examinees are divided into two groups, with one group taking the CBT and the other taking the PBT. The studies reviewed were fairly evenly divided between the two designs. Variations exist within each design, and these are discussed in the relevant section.

Within-subjects designs

In the within-subjects design, a single group of students is administered a PBT and a CBT. Counterbalancing is used to moderate any effects that might arise from test order (such as fatigue, practice, or motivation). With respect to presentation order, counterbalancing is a process by which examinees are divided into two groups: examinees in one group take the CBT first and the PBT second, and examinees in the other group take the tests in the opposite order. In addition, counterbalancing can be used to moderate other testing effects. For example, in the case of comparability, a within-subjects design often requires that examinees take one set of items on the PBT (e.g., form A) and another set of items on the CBT (e.g., form B). Thus, a researcher can counterbalance for both order and test form. Order can also be used as a blocking variable to reduce experimental error and to check whether one order leads to better performance than another. The within-subjects design is somewhat limited in its application because of the administrative problems caused by testing the same students twice. However, this design is the richest source of information concerning comparability. The only threat to the validity of this design is that taking two tests may not generalize to a population where only one test is administered.

There are three general variants of the within-subjects design. The first variant administers the same items (or draws items from the same testing pool) for the CBT and PBT. The second uses different items (i.e., test forms) for the CBT and PBT. The third variant does not administer a test twice to examinees. Rather, constructed responses are scored by raters in different formats (i.e., typed and handwritten). The studies representing each variation appear in Tables 2 through 4.

The first variant of within-subjects studies used the same set of items for the PBT and CBT. Table 2 summarizes details of these three studies. Key elements of the design of the studies are outlined in the bulleted text below.

- Pomplun and Custer (2005) administered the same items via a computer-based and paper-based test, counterbalancing the order. They administered the Initial Skills Analysis Test, which had three subtests (basic skills, comprehension, and language) to children in kindergarten to grade 3. Schools were randomly assigned to the counterbalanced condition. The purpose of the study was to conduct a confirmatory factor analysis on the raw subtest scores and mode to determine levels of factorial equivalence (factor structure equivalence, same factor loadings, same factor loadings and errors). Separate factor analyses were conducted at each of the four grades. A potential threat to the validity of this study is that

examinees were administered the same items twice, increasing the potential for a practice or memory effect. However, the random assignment and counterbalancing presumably minimized this threat.

- Eignor (1993) studied the performance of a group of high school juniors, high school seniors, and college freshmen in motivated conditions on computerized adaptive and paper-based versions of the SAT. The purpose of this study was to examine the degree of differences in equating scores from the two methods. The study design called for random counterbalancing the modes, but the instructions for counterbalancing were not closely followed. The item pool for the adaptive test shared items with the paper-based test, and presumably the adaptive test used item parameters calibrated from a previous paper-based administration, although this was not specified. The major threats to this design are that examinees may encounter the same items twice and thus increase the potential for a practice or memory effect, the problems with counter-balancing, and the potential confounds of using PBT-derived item parameters for the computer adaptive test.
- Olson, Maynes, Slawson, and Ho (1989) examined mode differences for CAT, CBT, and PBT in math for 3rd and 6th graders. The researchers counterbalanced the mode for the adaptive and computer-based test, and counterbalanced the mode for the adaptive and paper-based test. Students were randomly assigned into groups. The item pool for the adaptive test shared items with the paper-based and computer-based tests, and presumably the adaptive test used item parameters calibrated from a previous paper-based administration, although this was not specified in the report. The major threats to the validity of this design are similar to those listed for Eignor (1993).

Table 2. Studies Using a Within-Subjects Design (Counterbalancing for Order) in which the Same Items were Administered across Modes

Study	Study details
Pomplun and Custer (2005)	<i>Instruments:</i> Initial Skills Analysis Test (basic skills, comprehension, language), multiple choice w/reading passages <i>Sample:</i> Kindergarten (n=537), 1 st grade (n=457), 2 nd grade (n=498), 3 rd grade (n=467) <i>Other measures:</i> Parental income, indicated by whether examinee received free lunch <i>Purpose:</i> Score and construct equivalence
Eignor (1993)	<i>Instruments:</i> SAT-Verbal and Quantitative, multiple choice <i>Sample:</i> HS / College students (n=506) <i>Other measures:</i> None <i>Details:</i> CBT was an adaptive test <i>Purpose:</i> Score and construct equivalence
Olson, Maynes, Slawson, and Ho (1989)	<i>Instruments:</i> California Assessment Program math, multiple choice <i>Sample:</i> 3 rd grade (n=350), 6 th grade (n=225) <i>Other measures:</i> Testing time <i>Details:</i> Two CBTs: one adaptive, one linear <i>Purpose:</i> Score and construct equivalence

In the second variation of within-subjects studies, examinees are administered different items (i.e., different forms) across modes and these forms are counterbalanced as well. In one particularly interesting design, all examinees are administered the same test in the same mode, and then are administered both a computer-based and paper-based test using different items. The initial test is used for common-item equating, and permits equating studies to determine mode effects. Table 3 summarizes details of these studies, and key elements of the design are described in the bulleted text below.

- In Johnson and Green (2006), participants were randomly assigned into four groups in which counterbalancing occurred for testing order and form. Forms were not equated, and comparisons of scores were made within forms. The forms consisted of eight mathematics items, and students' processes in answering items were also captured by analyzing student worksheets and interviews about their process. Participants were 10-11 year old school children. Threats to the validity of this design were in the instrumentation (the instruments used were quite short and researcher-created) and in the small sample size (about 50 examinees per form).
- In Pomplun, Frey, Becker, and Hughes (2000), participants were randomly assigned into four groups where counterbalancing occurred for order and form. Six schools were randomly assigned form and mode; schools were asked to randomly assign students to test orders. The researchers administered the Nelson Denny Reading test at the high school and college levels. The focus of the study was the effect of mode on reading rate. However, different stopping criteria were used for measuring the reading rate. On CBT, the student clicked on the last word read and on the PBT, the student marked the last line read (and the middle word for that line was used). The threats to the validity of this design were the use of different stopping criteria, which could influence the calculation of test scores, and the small sample size.
- Poggio, Glasnapp, Yang, and Poggio (2005) conducted a comparability study on four already-equated forms from a state 7th grade mathematics assessment in Kansas. Forms were randomly assigned but the counterbalancing for mode order was not random; rather, volunteer schools chose which mode to administer first. The majority of schools chose to administer the CBT first. Because forms in either mode were randomly assigned, a subset of students took the same form in both modes. The major threat to the validity of this design is the non-randomized counterbalancing procedure because this could potentially results in non-equivalent groups. In addition, the sample size was small given the analyses conducted on these data.
- In Choi and Tinkler's (2002) study, participants in two randomly assigned groups took three tests. Participants first took a set of common items via computer, and then took either a computer-based test and then a paper-based test or a paper-based test and then a computer-based test. Random assignment into testing order was conducted at the classroom level. Two test forms were used. Both test mode and form were counterbalanced, although order differences were not presented. The instruments were state math and reading assessments and were administered to 3rd and 10th graders. The threat to the validity of this design is the unknown influence of the first CBT.

Table 3. Studies Using a Within-Subjects Design (With Counterbalancing for Order and Form) in which Different Items were Administered Across Modes

Study	Study details
Johnson and Green (2006)	<i>Instrument:</i> Mathematics test items aligned to the British National Curriculum, multiple choice <i>Sample:</i> British 10-11 year olds (n=104) <i>Other measures:</i> Observations of examinee test-taking, analysis of examinee test-taking strategies, interviews about mode preference <i>Purpose:</i> Score and construct equivalence
Pomplun, Frey, Becker, and Hughes (2000)	<i>Instrument:</i> Nelson Denny Reading Test (vocabulary, reading comprehension, and total score), multiple choice <i>Sample:</i> High school, 2 and 4 yr college (n=185) <i>Other measures:</i> None <i>Purpose:</i> Score and construct equivalence
Poggio, Glasnapp, Yang, and Poggio (2005)	<i>Instrument:</i> Kansas Computerized Assessment in mathematics multiple choice <i>Sample:</i> 7 th grade (n=646) <i>Other measures:</i> Gender, Socio-Economic Status (lunch support), and academic placement (general education, gifted, special education) <i>Purpose:</i> Score and construct equivalence
Choi and Tinkler (2002)	<i>Instruments:</i> Math and reading tests were portions of operational tests, multiple choice with stimulus material <i>Sample.</i> 3 rd grade (n~800), 10 th grade (n~800) <i>Other measures:</i> Additional CBT administered for equating, Analysis of reading item characteristics <i>Purpose:</i> Score and construct equivalence

In addition to the design above, a single-group within-subjects design has been used in studies whose focus was to determine whether *typed* constructed-response essays would be graded differently by scorers than *handwritten* essays. This approach could also be used to compare digitized handwritten essays and typed essays. In these studies, handwritten essays were transcribed into computerized text. Essays in both modes were then scored by human raters using a scoring rubric, and mean comparisons of the scores were calculated. These studies have been used in operational testing of elementary school students (Russell & Tao, 2004), middle school students (Hollenbeck, Tindal, Stieber, & Harniss, 1999; Russell & Tao, 2004), and high school students (Russell and Tao, 2004) in English Language Arts. In one study (Russell & Tao, 2004), raters also marked writing errors (e.g., spelling, punctuation, capitalization, awkward transitions, confusing passage) and compared the proportion of errors across mode. These studies involved no additional participation on the part of the student, since the student essays were re-scored. In Russell and Tao (2004), essays were transcribed twice into text as a single-spaced essay and as a double-spaced essay. The only threat to the internal validity of these types of studies is that no comparative group existed to separate rater reliability and bias from mode effects. Interestingly, none of these studies calculated rater agreement across mode. These studies also used small samples. Table 4 summarizes details of these studies.

Table 4. Special Studies Using a Within-Subjects Design (Same Responses, Different Raters)

Study	Study details
Hollenbeck, Tindal, Stieber, and Harniss (1999)	<i>Instrument:</i> Single Oregon ELA writing essay item <i>Sample:</i> Middle school students (n=80) <i>Other measures:</i> None <i>Purpose:</i> Score equivalence
Russell and Tao (2004)	<i>Instrument:</i> Massachusetts Comprehensive Assessment System (MCAS) Language Arts Test, essay <i>Sample:</i> 4 th grade (n=52), 8 th grade (n=60), 10 th grade (n=60) <i>Other measures:</i> Analysis of essay features <i>Purpose:</i> Score equivalence

Between-subjects designs

In the between-subjects design, participants are divided into two or more groups and each group is administered the same items via a computer-based test or a paper-based test. There are no serious threats to the internal validity of these designs if the groups are assumed to be randomly equivalent.

Three variations of the between-subjects design are found in the studies. The first variation uses random or pseudo-random methods to assign groups. The second variation uses matching methods to form groups. Finally, the third variation uses additional test data to perform post-hoc covariance analysis or for equating. Tables 5 through 7 provide information on these studies.

In the first variation of between-subjects studies, examinees are divided into groups using random or pseudo-random methods (e.g., stratified random sampling). There are no general threats to the validity of this type of design. Table 5 summarizes details of these studies, and they studies are described in the bulleted list below.

- A series of studies on the Texas statewide graduation exams (TAKS) in reading, mathematics, social studies and science (Way, Davis, & Fitzpatrick, 2006; Way & Fitzpatrick, 2006; Keng, McClarty, & Davis, 2006) used the between-subjects approach. Eleventh graders who had failed the graduation exam were offered an additional testing opportunity, and were randomly assigned to take a PBT or CBT. The major threats to the validity of this design are the restricted range of scores due to using a sample that had scored relatively low on the original administration and problems of generalizing of the re-test results to another population.
- Fitzpatrick and Triscari (2005) randomly assigned subjects to two groups (CBT or PBT) in order to examine the comparability of the operational Virginia End of Course Algebra I, Earth Science, and English Language Arts tests. Scores from a previously-taken PBT test were used to examine the similarity of the two groups, and the groups were found to be too

different to compare scores. As a result the study used a non-equivalent groups, common-item linking design to examine equating parameters.

- Stratified random samples are also used to divide examinees into groups. Russell and Plati (2001) used English grades from a previous year to assign participants to a CBT or a PBT group in 8th and 10th grade samples. In this study, responses to a single extended response item were examined to determine whether writing an essay on computer differed from writing an essay on paper. All handwritten essays were transcribed and subsequently scored by two raters blind to the testing mode. The raters' scores were then summed. The authors indicated that the motivation levels may have differed between the two modes due to the tests being administered in a low motivation condition and the use of different proctors. Also, a single high performing district was used. In another study with a similar design, Russell (1999) used the grade 7 Stanford Achievement Test 9 (SAT-9) normal curve equivalents (NCE) to assign examinees for an 8th grade sample in math, language arts, and science. This study focused on the influence of mode on constructed-response items. The major threats to the validity of these designs were the limited sample, the instrumentation consisting of few items, and the potential differences in motivation.
- Various computer-based test characteristics were studied (e.g., enabling examinees to scroll line-by-line through long text passages or to scroll page-by-page). Researchers (Higgins, Russell, & Hoffman, 2005; Pommerich, 2004) divided examinees randomly into various computer-based test conditions and one paper condition. Pommerich (2004) conducted two studies in an effort to examine the impact of a CBT interface. One examined mode differences between PBT and CBT scores. The second study used results learned in the first study, and examined the influence of various item presentation modes in CBTs. In this study, two forms of automated scrolling were used in English. In one variation, the relevant portion of the passage automatically scrolled if it did not appear in the passage window. In the second variation, the relevant portion of the passage automatically scrolled to the top of the page for each item. The influence of line-by-line and page-by-page scrolling was examined in reading and science reasoning. The other examined the influence of two methods for automated scrolling through stimulus passages. Higgins, Russell, and Hoffman (2005) examined the impact of item presentation on paper, on computer with line-by-line scrolling, and on computer with page-by-page scrolling. The authors used a 4th grade reading test with long passages in their study. In addition, the authors examined results in relation to external measures such as a computer fluidity test.

Table 5. Studies Using a Between-Subjects Design with Random or Pseudo-Random Assignment

Study	Study details
Way and Fitzpatrick (2006); Way, Davis, and Fitzpatrick (2006); Keng, McClarty, and Davis (2006)	<i>Instrument:</i> Texas Assessment of Knowledge and Skills (TAKS) Math, ELA, Social Studies, Science; multiple choice. <i>Sample:</i> 11 th grade retest. Math (n=2156), Science (n=2201), Social Studies (n=743), ELA (n=1368) <i>Other measures:</i> Survey administered regarding computer skills, use, and preference for testing <i>Purpose:</i> Score and construct equivalence

Study	Study details
Fitzpatrick and Triscari (2005)	<i>Instrument:</i> Virginia state end-of-course tests in Algebra, Earth Science and ELA, multiple choice <i>Sample:</i> High school students (n=2205) <i>Other measures:</i> Additional PBT administered prior to study <i>Purpose:</i> Score equivalence
Russell and Plati (2001)	<i>Instrument:</i> Single Massachusetts Comprehensive Assessment System (MCAS) writing prompt <i>Sample:</i> 8 th grade (n=144) and 10 th grade (n=145) <i>Other measures:</i> Keyboarding test, survey of prior computer use, mid-term grade used as covariate <i>Purpose:</i> Score and construct equivalence
Russell (1999)	<i>Instrument:</i> Items from NAEP and Massachusetts Comprehensive Assessment System (MCAS); math, science, and ELA, all items were constructed-response <i>Sample:</i> 8 th grade (n=229) <i>Other measures:</i> Keyboarding test, computer use survey, and SAT-9 NCE scores used to stratify sample <i>Purpose:</i> Score and construct equivalence
Higgins, Russell, and Hoffman (2005)	<i>Instrument:</i> Reading Comprehension Test, multiple choice with passages. Released items from NAEP, Progress in International Reading Literacy Study (PIRLS), and NH state assessments <i>Sample:</i> 4 th grade (n=219) <i>Other measures:</i> Computer fluidity test, computer literacy test, computer use survey <i>Purpose:</i> Score and construct equivalence
Pommerich (2004)	<i>Instrument:</i> English, Reading, and Science Reasoning content areas. Test original unknown. Most items had stimulus materials. All items were multiple choice. <i>Sample:</i> 11 th and 12 th grade (Study 1: n= 5612, Study 2: n= 9473) <i>Other measures:</i> None <i>Purpose:</i> Score and construct equivalence

The second variation of the between-subjects design used groups which took the tests at different times. In these designs, the CBT group was a volunteer sample, and the PBT group was assembled from prior testing. In some studies this group is the entire census tested group or a subset of the entire group. There are a number of threats to the validity of this type of design. First, if the administration of the CBT is not comparable to that of the PBT, then the data will likely not be comparable. Second, maturation or additional learning may have occurred if the CBT testing occurs much later than the PBT testing. Third, the sample of students who agree to participate in the CBT testing may not have the same characteristics as the students participating in the PBT testing. Table 6 summarizes details of these studies.

- Schwarz, Rich, and Podrabsky (2003) examined the mode comparability of the analytical reasoning and quantitative reasoning subscales of InView Test. The CBT sample data were collected following the much larger PBT standardization sample. The CBT sample was selected based upon region city size, and lunch status. The authors did not provide information on the method for selecting participants. A threat to this design is that the researchers used the entire norming sample rather than a subsample selected to be equivalent to the examinees participating in the study.
- In Way, Davis, and Fitzpatrick's (2006) study of the TAKS, the small number of study volunteers did not allow for enough participants to randomly assign subjects to CBT and PBT conditions. The investigators administered a CBT to all participants, and used a set of matched samples from all students taking the operational PBT tests. Eighth graders were the subjects of study, and the 7th grade TAKS reading and math test scores were used as the matching variables.
- In a National Assessment of Educational Progress (NAEP) study, Sandine, Horkay, Bennett, Allen, Braswell, Kaplan, and Oranje (2005) used this design to study mode effects of writing on a computer and on paper. The study data were collected in the months immediately following the operational NAEP administration. A random sample of schools participating in NAEP was chosen for the study, and students with existing NAEP reading or writing data were selected to be in the sample. One half of the sample had taken a NAEP writing item, and one half had taken the NAEP reading (and not the writing item) item. The latter sample was collected to examine whether taking a NAEP writing item previously influenced performance; it did not. Study participants answered two constructed-response prompts on the computer. The CBT scores were compared to their PBT scores, and to a comparable sample of respondents answering the same essay during the operational NAEP assessment. Potential threats to this design are that the sample did not reflect the nationally representative NAEP sample, and that only two items were used in the instrument.

Table 6. Studies Using a Between-Subjects Design with Volunteer CBT Sample and Pre-Existing Sample

Study	Details
Schwarz, Rich, and Podrabsky (2003)	<p><i>Instrument:</i> InView Analytical Reasoning and Quantitative Reasoning subtests, selected response (i.e., multiple choice, matching)</p> <p><i>Sample:</i> Analytical Reasoning: Grades 4 and 5 (n=2295), 6 and 7 (n=1839), 8 and 9 (n=1455); Quantitative Reasoning Grades 4 and 5 (n=2103), 6 and 7 (n=1623), 8 and 9 (n=1260)</p> <p><i>Other measures:</i> None</p> <p><i>Purpose:</i> Score and construct equivalence</p>
Way and Fitzpatrick (2006); Way, Davis, and Fitzpatrick (2006); Keng, McClarty and Davis (2006)	<p><i>Instrument:</i> TAKS Reading, Math, Social Studies, multiple choice and constructed response</p> <p><i>Sample:</i> Grade 8. Math (n=1273), Social Studies (n=1449), Reading (n=1840)</p> <p><i>Other measures:</i> None</p> <p><i>Purpose:</i> Score equivalence</p>

Study	Details
Sandine, Horkay, Bennett, Allen, Braswell, Kaplan, and Oranje (2005)	<p><i>Instrument:</i> Two essays used in the main NAEP Writing assessment</p> <p><i>Sample:</i> 8th grade (n=1255)</p> <p><i>Other measures:</i> Background questions that asked for demographic information, computer experience, and writing instruction. Also, an online computer skills measure.</p> <p><i>Purpose:</i> Score and construct equivalence</p>

In the third and final variation of the between-subjects design, examinees take additional tests aside from the tests being examined for comparability. These additional tests are used to equate the item parameters and test scores, or the test scores are used as covariates. The additional tests are often administered prior to the studied tests. Collecting the additional test data enables the measurement of group differences, the use of regression methods to control for differences, and the information for equating the PBT and CBT. The threat to these designs is the potential influence of the initial PBT which is often administered first and is not counterbalanced during the experiment. Taking the initial PBT may produce practice or fatigue effects that are not controlled through counterbalancing. Table 7 summarizes the details of these studies and the bulleted list below provides additional information.

- Russell and Haney (1997) randomly drew examinees from all students enrolled in a special school. All examinees were administered a set of PBT open-ended items in writing, science, math, and reading. Examinees then took a set of NAEP items (science, math, and language arts) and one extended writing item either in a computer-based or paper-based format. The open-ended items taken on paper by all examinees were scored by a single rater. For the two tests examined for mode differences, the hand-written responses to open-ended items were typed verbatim into the computer to minimize rater bias. These items were scored by three raters, and the average score across the three raters was used as the dependent variable. The groups were found to differ on the open-ended items taken on paper, so these scores were used as a covariate in analyzing the NAEP items and writing item. Speededness was also examined across the two groups, and the two groups were found to differ in their test completion status, with the CBT examinees having higher test completion rates. The researchers noted that the proctors may have allowed extra time to students in the CBT condition because it was a new experience. Threats to the validity of this design are the use of an instrument having potentially significant error (the open-ended items) when using covariance methods, and the possibility of differences between test conditions.
- In a NAEP study (Sandine et al., 2005), researchers administered a block of twenty items on paper all examinees, and then administered either a computer-based or paper-based test. All examinees answered a set of background questions at the end of the test administration session. The subject under study was math, and participants were 8th grade students. A multi-stage, probability-based sampling strategy was used. Open-ended responses were scored in the mode in which they were taken. The paper-administered common items were used both as a covariate for mean comparisons and to enable item-level comparisons. Item response theory was used to estimate proficiency values, which were then transformed into scale

values. Aside from the threats described above, there were no threats to the validity of this design.

- Zhang and Lau (2006) used SAT-9 scores as the common test, and then a used state test for their comparability study in reading (5th and 8th graders) and math (8th graders). The students in the study were those requiring a retest, having been assigned into “below standard” and “well below standard” from the first test. The common test was used to equate the item parameters and create RS-SS tables, and then compare the RS-SS tables for the two modes. Students were not randomly assigned into conditions as assignment into a condition was based upon consultation with the student. Thus, the threats to the validity of this design are potential group differences due to self-selection into groups, as well as restriction of range and issues with generalizability due to the use of re-test students.

Table 7. Studies Using a Between-Subjects Design and Additional Tests

Study	Details
Russell and Haney (1997)	<p><i>Instruments:</i> Sub-test of NAEP multiple choice and constructed response items in language arts, science, and math. Locally-created performance writing item. Locally-created sub-test using all open-ended items in writing, science, math, and reading (used as a covariate).</p> <p><i>Sample:</i> 6th, 7th, and 8th grade (n=89)</p> <p><i>Other measures:</i> Locally-created sub-test (administered on paper) used as a covariate</p> <p><i>Purpose:</i> Score equivalence</p>
Sandine, Horkay, Bennett, Allen, Braswell, Kaplan, and Oranje (2005)	<p><i>Instruments:</i> NAEP math items, multiple choice and constructed response</p> <p><i>Sample:</i> 8th grade (n=1970)</p> <p><i>Other measures:</i> Additional PBT administered as a covariate, independent review of items for extent of modification for CBT, Background survey asking about demographic variables and computer use, familiarity, and skills.</p> <p><i>Purpose:</i> Score and construct equivalence</p>
Zhang and Lau (2006)	<p><i>Instruments:</i> Delaware state assessment, SAT-9 Reading and Math, mostly multiple choice with some short-answer and extended-response items</p> <p><i>Sample:</i> 5th grade (n=570) and 8th grade (n=330) reading, 8th grade math (n=801)</p> <p><i>Other measures:</i> Additional PBT administered and used for common-item equating, survey of test use, observations of testing conditions</p> <p><i>Purpose:</i> Score equivalence</p>

Matching designs

Matching designs are a viable alternative when random assignment and/or repeated measures studies are not possible. In matching designs for examining mode comparability, a small study

sample is administered a test in one mode (often, a CBT) and a comparable sample based on key matching variables is drawn from a larger sample administered a test in another mode (often, a PBT). The small study sample is likely a special sample of volunteers who take a test on computer. The larger PBT sample is likely gathered from operational PBT testing. A primary goal in matching designs is that the matched sample is comparable on the critical variables of concern. Two promising approaches to creating a matched sample are described below: the Matched Samples Comparability Analysis (MSCA) and Propensity Score Matching.

Matched Samples Comparability Analysis (MSCA)

Way, Davis, and Fitzpatrick (2006) introduced a matching design to conduct a comparability analyses in the context of linking. They called their design “Matched Samples Comparability Analysis (MSCA).” In their study, a small sample of CBT responses was collected and compared to a much larger sample of operational PBT responses. The MSCA uses a bootstrap design (i.e., random draws with replacement) with equating to determine the magnitude of equating error across modes and at all score points on a test. The MSCA design for this study used the previous year’s test scores as matching variable to conduct studies each of math, science, and social studies. Presumably, any combination of matching variables can be used.

The MSCA involves implementing the following procedure: 1) Collect a bootstrap sample from the CBT sample equal to the size of the CBT sample; 2) Collect a bootstrap sample of the same size from the PBT sample and matched on both matching variables; 3) Conduct raw score-to-raw score equating using IRT true score equating; and, 4) Transform the raw scores to scale scores using the operational RS-SS tables. The authors used 500 bootstrap replications in their study, although presumably any reasonable number of replications can be used. Once all samples have been collected, the CBT RS-SS tables can be created by averaging the scale scores across all samples at each score point. In addition, the standard deviation at each score point can also be calculated and reflects the error in linking at each raw score point.

In the same publication, the authors reported the results of a simulation study that examined two situations when examinees differ on ability based on prior test performance: a) the performance of MSCA when mode differences do not exist; and, b) the performance of MSCA to when mode differences do exist. Using PBT data in math and reading, the authors generated six datasets each with different frequency distributions. They examined the performance of MSCA in the context of mode effect differences of 0, .25, .5, and 1.0. The MSCA appropriately did not detect significance using a 95% confidence interval when no mode effects exist, and did detect significance when mode effect differences were .5 or 1. The MSCA had difficulty detecting significance when the mode effect was .25.

The MSCA approach has so far been limited to use in the TAKS testing program. Potential threats to the validity of this design are improper use of a matching variable (resulting in non-comparable groups on key variables) and differences in testing conditions such as low motivation or Hawthorne effects.

Propensity score matching designs

Propensity score matching (D'Agostino, 1998; Rosenbaum, 1995; Rosenbaum & Rubin, 1983; Rubin, 2006) is a relatively new and efficient method for producing a matched group design. The method is a refinement of a more general matching or covariate design. Matched groups are created to reduce bias that may result in a two-group design where randomization is difficult or impossible to implement fully, or when a within-subjects design is ruled out by administrative difficulties. Three examples where propensity score matching has been used are listed below.

- A study of the effect of offering accommodations, such as additional testing time, on student performance (Rudner & Peyton, 2006);
- The effect of taking a test before versus after graduation (Rudner & Peyton, 2006);
- The effect of a change in contractors and item pools on results of a computer-based test. There is currently no publicly available documentation for this example but, like the previous two examples, the study occurred recently with the Graduate Management Admissions Test (Schulz, M., personal communication, March 1, 2007).

In each of these examples, the experimental group was a relatively small, non-randomly selected group, and the control group was selected from a much larger population that experienced the control treatment. Generally, in a matching design, one selects, for each experimental subject, a control subject with the same standing on one or more matching variables. If only one variable or covariate is used for matching, there is essentially no difference between propensity score matching and traditional matching or covariate analysis. When two or more variables are used for matching, propensity score matching is more powerful and effective. The propensity matching technique uses multivariate logistic regression to form a weighted composite of the covariates—the propensity score—for matching. Propensity score matching is more effective because it uses a weighted composite of covariates, and because it can accommodate missing data and therefore include more covariates and even subjects.

In a comparability study, propensity score matching would work as follows: 1) a group of students is tested using a CBT, and background variables are measured or taken from existing student records; 2) an existing group of students (possibly quite large in number) which has already taken the PBT version of the test is assembled, complete with PBT score and background variables; 3) logistic regression is used to predict group membership, and each member of both groups is assigned a value of the likelihood of belonging to each group (i.e., a propensity score); 4) for each examinee who took the CBT, an examinee who took the PBT with the nearest propensity score is selected. Analysis of the CBT and PBT scores follow the matching. As with the MSCA approach, potential threats to the validity of this design are improper implementation of the match (resulting in non-comparable groups) and differences in testing conditions.

Types of analyses

In order to be considered comparable, the CBT and PBT should measure the same construct and should measure these constructs with the same degree of precision. The types of analyses used in comparability studies can be thought of in terms of the types of hypotheses tested. For example,

it was mentioned earlier that if the CBT and PBT measure the same construct, then the following should be true:

- The test content and content specifications must be the same.
- The scores should have the same factor structure.
- The scores should have the same measurement precision.
- The score distributions should differ only in difficulty, and hence, be equitable.
- The scores should be highly related to one another.
- The scores should have the same relationship to other related measures.

The comparability studies included in this review were classified above in terms of the experimental design employed. They are now classified in terms of the six content validity hypotheses (listed directly above) that were examined using the experimental data gathered by the chosen research design.

Test content

Content comparability between the paper-based and computer-based modes can be conceptualized as requiring that tests have the same test specifications, have similar items measuring each construct, and require the same skills to answer those items. This issue can be studied regardless of the study design chosen as it relates more to instrumentation than design.

Test specifications. In the comparability studies considered for this review, the computer-based and paper-based tests were built using the same test specifications. An exception to this rule is computer adaptive testing, where the administered tests vary for examinees, although presumably conditions are placed on the item pool and test administration to ensure parity. The items administered via the computer-based test were often transferred to the computer directly from paper-based administrations, so essentially the “same” items were administered in each mode.

Item similarity. The degree to which the items are similar across modes is related to the extent to which items need modification. Most studies simply described the process for transferring the paper-based items to the computer format. One method for ensuring similarity was that the same font and style are used (Higgins, Russell, & Hoffman, 2005), or that participants received a paper copy of the test booklet (Russell, 1999). One study had students use the same computers so there was no variation of resolution or screen size across CBTs (Higgins, Russell, & Hoffman, 2005). Sandine et al. (2005) examined the results for participants using study-provided laptops versus school computers. In addition, Sandine et al. (2005) also used an external review to characterize the amount of adaptation needed to use the item on the computer and used this characterization in its analysis of differences in item difficulty across mode. Another issue influencing item similarity is the use of online toolkits (such as a compass, calculator, or ruler). Additionally, the use of color as a navigational or attentional aid may have a differential effect for some examinees (such as color-blind students).

Skills required. Taking a test on computer may require skills unrelated to the construct. The studies attempted to reduce the influence of potential construct-irrelevant skills by attending to

potential differences in the test-taking process. The types of interface issues for the CBTs are listed below.

- Use of a tutorial (Higgins, Russell, & Hoffman, 2005; Sandine et al., 2005; Poggio, Glasnapp, Yang, & Poggio, 2005).
- Enabling item review throughout the test (Schwarz, Rich, & Podrabsky, 2004; Higgins, Russell, & Hoffman, 2005; Poggio, Glasnapp, Yang, & Poggio, 2005) or sections of the test (Pommerich, 2004) to better mimic a paper-based test.
- Allowing only one item per screen to be presented (Schwarz, Rich, & Podrabsky, 2004; Poggio, Glasnapp, Yang, & Poggio, 2005), or allowing multiple items to be presented for a testlet (Pommerich, 2004).
- Enabling or disabling spell-checking, grammar, and copy/paste functionalities (Russell & Plati, 2001; Sandine et al, 2005; Zhang & Lau, 2006).
- Using scrolling or whole-page advancement for text passages that do not fit on a single screen (Pommerich, 2004; Higgins, Russell, & Hoffman, 2005).
- Using highlighting or automatic scrolling through stimulus text as an attentional aid (Higgins, Russell, & Hoffman, 2005; Pommerich, 2004).

Related to the issue of skills required, one study (Zhang & Lau, 2006) also reported on the importance of ensuring that staff involved with the CBT administration was properly trained.

Factor structure

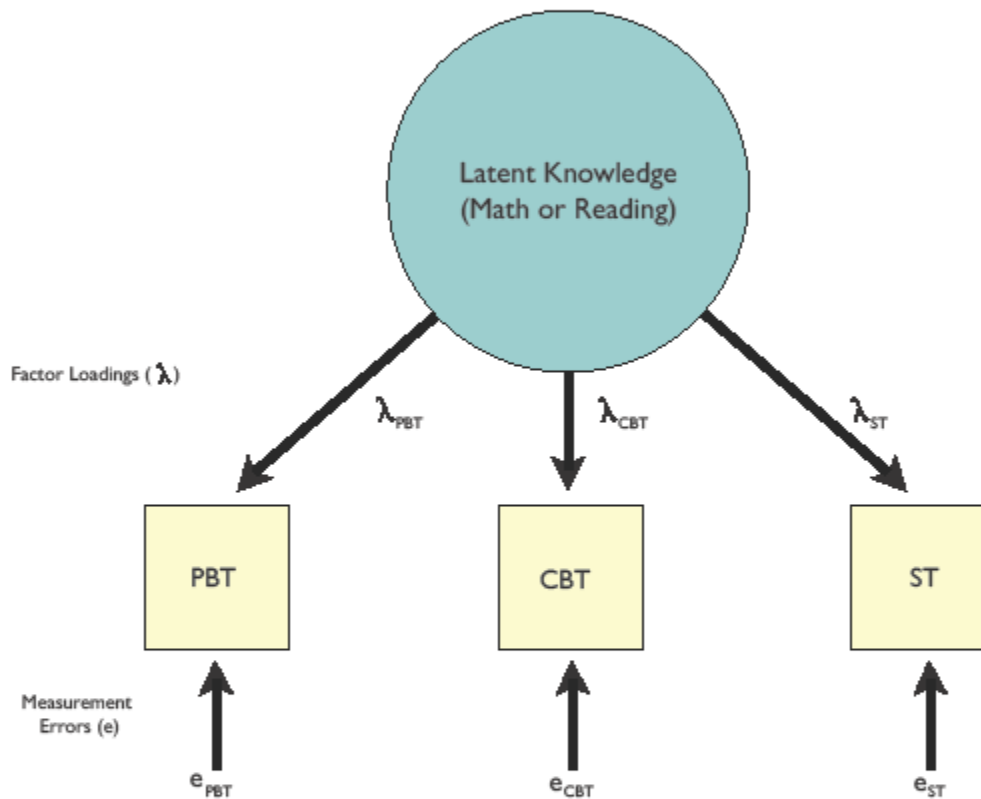
If two instruments measure the same construct, then they should have the same factor structure. In comparability studies, this hypothesis can be examined by using confirmatory factor analysis techniques or a full structural equation model to examine the fit of various hypothesized models. This type of study can only be examined using a within-subjects design because data from paper-based and computer-based tests using the same items need to be collected from each participant. Figure 1 provides a visual depiction of a confirmatory factor analysis approach that might be used. Only one study (Pomplun & Custer, 2005) from the K-12 comparability literature examined the invariance of factor structure across mode. This technique has been used in comparability studies on adults (Boo, 1997; Neumann & Baydoun, 1998).

The Pomplun and Custer (2005) study used a confirmatory factor analysis to determine whether a CBT with three subtests differed in its factor structure from the PBT. Three models were tested. All models used a three-factor model with subtests from each model loading on the appropriate factor. The three factors were allowed to covary, and no error covariances were allowed among the observed variables. The first model placed no constraints on the factor loadings or error variances, and therefore just looked at pattern consistency (i.e., congenericity). The second model constrained the factor loadings to be the same within a factor, and thus required the modes to have the same relationship to the factor (i.e., tau equivalence). The third model constrained the factor loadings and error variances to be the same within a factor, and thus required the modes to have the same relationship to the factor and same error (i.e., parallelism). Model fit was tested using statistical tests and fit indices.

Although not a K-12 study, Moreno and Segall's (1997) structural equation modeling approach on Armed Services Vocational Aptitude Battery (ASVAB) data bears noting. In this study,

participants with a prior PBT test score were randomly assigned to two groups. In one group, participants took two additional forms of the PBT. In the other group, participants took two computer-adaptive tests (CAT-ASVABs) built to the same specifications as the PBT. Instead of testing model fit for parallelism, tau-equivalence, or congeneric equivalence, these researchers made assumptions regarding relationships among PBT scores and CBT scores (assuming latent correlations among scores were 1 within mode) and estimated the test reliabilities and the relationship between the CBT and PBT. This analysis was conducted for ten subtests of the ASVAB. The researchers then compared the estimated reliability among tests and presented the disattenuated correlations between PBTs and CBTs (all of which were close to 1). Significance tests for model fit supported the hypotheses that the CBT-PBT correlation was the same as the PBT-PBT and CBT-CBT correlations.

**Confirmatory Factor Analysis Model for
Examining the Comparability of
Paper-Based and Computer-Based Tests (PBT_s and CBT_s)
[Using an Additional Summative Test (ST)]**



Hypotheses	Statistical Tests
1) PBT, CBT and ST have perfectly correlated True Score	Fit of overall model
2) PBT and CBT are parallel	$\lambda_{PBT} = \lambda_{CBT}$ and $\sigma^2(e_{PBT}) = \sigma^2(e_{CBT})$
3) PBT and CBT have same validity for predicting ST	$\lambda_{PBT} = \lambda_{CBT}$

Figure 1. Confirmatory Factor Analysis Model Used for Assessing Comparability of a PBT and CBT

Item-level differences can also be considered an indication of factor structure. Item differences were examined using a variety of methods. Once item-level differences were determined, the items were reviewed by researchers to explain why such differences occurred.

- Schwarz, Rich, and Podrabsky (2003) used differential item functioning using PBT as the reference group and CBT as the focal group. Poggio, Glassnap, Yang, and Poggio (2005) conducted item-level and category-level DIF studies with small sample sizes.
- Pommerich (2004) and Higgins, Russell, and Hoffman (2005) used confidence intervals to determine whether item p -values varied significantly across mode. Pommerich (2004) presented confidence intervals for each item studied, and also presented the proportion of items statistically favoring each mode.
- Sandine et al. (2005) compared item p -values and IRT b parameters, as well as presented scatterplots of a and b parameters across mode. The comparison of parameters was also categorized by the amount of modification needed to computerize the items. This study had a set of common items administered in the same mode, and these item parameters were constrained to be equal in the calibration process, whereas the parameters of items administered in the two modes were allowed to vary.
- Choi and Tinkler (2005) used a similar methodology as Sandine et al. (2005). Choi and Tinkler (2005) also categorized items by the level of comprehension required (comprehension at the word, phrase, sentence, or discourse level) to answer the item, and then examined bias and error between the b values by mode for each category.
- Johnson and Green (2006) compared p -values of items as well as conducted a qualitative examination of error types (e.g., transcription error, place value error, partial answer, computation error, misunderstanding) made by students in each mode as well as strategies across modes.
- Keng, McClarty, and Davis (2006) examined differences in mode at the item level by comparing p -values, differences in choices across response options, and computing IRT-based differential item functioning tests. These researchers also reviewed a sample of test booklets to determine whether scratchwork on the booklet was associated with predicted item performance.

Measurement precision

If two tests are comparable, then they should have the same measurement precision both overall and across proficiency levels. This issue can be studied using any design presented, although the indicator of measurement precision may vary across designs. Measurement precision can be examined at the overall test level and can be examined by looking at the consistency of individual constructed-response item ratings.

Two methods for examining measurement precision at the test level are Cronbach's alpha and IRT information curves. Cronbach's alpha and the standard errors of measurement should be the same across modes, as should the test information functions. Classical test theory reliability was estimated in a few studies (Olson, Maynes, Slawson, & Ho, 1989; Zhang & Lau, 2006). Surprisingly, only one study provided IRT test information curves to provide a comparison of standard error conditional on theta for modes (Poggio, Glassnap, Yang, & Poggio, 2005).

Rater agreement statistics (such as Cohen's kappa, exact agreement rates, and the proportion of essays needing adjudication by an external rater) can be used to compare consistency for constructed-response items. While researchers often provided indices of measurement precision across modes, few provided indices within modes and then compared precision across modes. Sandine et al. (2005) provided exact agreement rates for constructed-response items for both modes in the NAEP assessments. Way and Fitzpatrick (2006) compared rater agreement (kappa, exact agreement, and proportion needing adjudication by a third rater) for a single essay by mode for the TAKS. This study also examined the impact of mode using automated essay scoring. Five samples (71 in human scorer calibration, 300 PBT, 300 CBT, 300 PBT and 300 CBT, 150 PBT and 150 CBT) were used to calibrate the automated essay scoring engine, and then the five trained engines were used to score both handwritten and computer-entered essays. Mode effects were detected when the engine trained in one mode scored same-mode essays more consistently than other-mode essays.

Score distributions

If two tests are comparable across modes, then their score distributions (e.g., means, standard deviations, frequency distributions) should be the same. A comparison of score distributions can be made for any of the designs described. Equating studies and the computation of RS-SS tables also provide information about score distribution.

Score comparisons. Comparisons of mean test scores were made in all the studies considered. Studies compared raw score means (Hollenbeck, Tindal, Stieber, & Harniss, 1999; Russell & Tao, 2004; Schwarz, Rich, & Podrabsky, 2004; Way, Davis, & Fitzpatrick, 2006; Russell, 1999; Sandine et al., 2005; Russell & Haney, 1997; Zhang & Lau, 2006; Eignor, 1993; Johnson & Green, 2006; Olson, Maynes, Slawson & Ho, 1989), scale score means (Schwarz, Rich, & Podrabsky, 2004; Way, Davis, & Fitzpatrick, 2006; Zhang & Lau, 2006; Pomplun & Custer, 2005; Pomplun, Frey, Becker, & Hughes, 2000; Poggio, Glasnapp, Yang, & Poggio, 2005), and IRT proficiency (theta) score means (Choi & Tinkler, 2002; Sandine et al., 2005). For studies using IRT, item parameters were constrained to be equal across mode so that any mode differences appeared in the proficiency scores. Other measures of distributional differences were computed in only a few studies. Pomplun and Custer (2005) examined equivalence of variance of mean scores. Eignor (1993) presented histograms of scores. When groups differed on the pre-test, researchers often used multiple regression to obtain means adjusted for those differences (Russell & Haney, 1997; Russell, 1999).

Equating. Researchers conducted equating studies to examine mode effects using a variety of different methods.

- Choi and Tinkler (2002) calibrated items in an IRT program twice, once indicating group membership and once not indicating membership. They then used a statistical test to compare model fit. A test administered to the two groups in the same mode was used as a basis for comparison.
- Zhang and Lau (2006) examined the mode effect using a PBT test administered to all examinees prior to the test examined for comparability. This PBT was used for common-item equating and so differences in equated parameters could be examined across mode.

- Fitzpatrick and Triscari (2005) equated the item parameters from each mode to existing Rasch item parameters gathered from an operational paper-based test administration. A difference between equating parameters would then be attributed to mode differences (although no overall difference was detected).
- Way, Davis, and Fitzpatrick (2006) used a bootstrap technique to determine standard errors in equating paper-based and computer-based test scores. Regions of the IRT proficiency scale were noted where scale score differences were outside the 95% confidence interval. The researchers also conducted a simulation study to determine whether this bootstrap methodology would be able to distinguish between differences in mode and differences in group ability. This study examined the standard errors produced when no mode effects existed but group effects existed, and when mode effects existed but group effects did not.

RS-SS tables. Closely related to equating was the comparison of raw score to scale score tables. Three studies computed RS-SS tables following an equating procedure and then compared the scale scores across the raw score range between the two modes (Way, Davis, & Fitzpatrick, 2006; Fitzpatrick & Triscari, 2005; Eignor, 1993). Because Way, Davis, and Fitzpatrick (2006) had computed standard errors in the equating process, they could determine where the scale scores differed significantly across the two modes. Eignor (1993) conducted both linear and curvilinear (equipercentile) equating to determine which method was appropriate and then computed RS-SS tables for each mode and testing order. In one test, the RS-SS tables were similar within mode and across order, so those data were combined. On another test, the RS-SS tables differed within mode and across order. As a result, these results were combined using a weighted sum. Fitzpatrick and Triscari (2005) generated separate RS-SS tables for the CBT and PBT, applied cut-scores derived for the PBT score to the CBT scores, and then examined the proportions of PBT and CBT examinees scoring at each proficiency level. Two studies (Zhang & Lau, 2006; Choi & Tinkler, 2002) presented differences in test characteristic curves across the proficiency continuum to reflect mode differences.

Relationship of scores

If the two tests are comparable, scores gathered from both tests should be highly related. If norm-referenced comparisons are being made, then the correlation corrected, for unreliability, should be 1.¹ If criterion-referenced judgments are made, then the agreement among levels should be very high (80% or higher agreement). Relationships among scores can only be computed in within-subjects designs (and possibly in matched-group designs). Correlations of raw, scaled, and/or IRT proficiency scores were computed in a number of studies (Sandine et al., 2005; Eignor, 1993; Pomplun & Custer, 2005; Olson, Maynes, Slawson, & Ho, 1989; Pomplun, Frey, Becker, & Hughes, 2000; Poggio, Glasnapp, Yang, & Poggio, 2005). None of the within-subjects designs reported a statistic on rater agreement.

The analysis of assignment of examinees into performance categories is an important issue that was not often considered in the studies. Analyses of performance category placement are

¹ A note of caution is in order here: if coefficient alpha is used in correcting for unreliability, this will often result in an over-correction because alpha is a lower-bound estimate to reliability.

compelling because these placements are of primary concern to many K-12 stakeholders. Only one study (Zhang & Lau, 2006) compared the proportion of examinees assigned into performance levels by mode. This study also presented agreement tables of performance level assignment from each mode of the studied test and from a test administered a few months prior. Thus, assignment into categories could be compared across mode. Court (2006)² conducted a follow-up study to Poggio, Glasnapp, Yang, & Poggio (2005) using separately collected data from a within-subjects design. This study presented the overall proportion of examinees assigned to performance levels by mode as well as the proportion of examinees being scored in the same or a different category based upon the mode of test.

Relationship with other variables

If two tests are comparable, then they should relate to other factors to the same degree. These factors may be specifically related to the test under study (such as completion rate, time spent) or related to the construct under study.

Test issues. Researchers have examined a variety of factors related to the test itself. The list of studied characteristics is listed below, along with the studies in which these factors were examined.

- Completion rate (Higgins, Russell, & Hoffman, 2005; Pommerich, 2004; Sandine et al., 2005; Russell & Haney, 1997)
- Time to complete the test (Russell & Plati, 2004).
- Surface features of constructed-response text such as number of characters, number of words, variation in sentence length (Russell & Plati, 2004; Way & Fitzpatrick, 2006; Russell and Haney, 1997; Sandine et al., 2005).
- Proportion of valid responses to constructed-response items (Sandine et al., 2005).

Other factors. The relationship of the test to other constructs has also been studied. This work has generally been done using regression or ANOVA. In these models, the impact of mode was examined controlling for other factors (such as a pre-test, typing speed) or interactions were examined for other factors (such as gender and race). One problem with this analysis approach is that regression methods assume the variables have no measurement error, and the influence of error on weights is unknown. Structural equation modeling is one methodology that could be used to account for error in the measure and compare relationships among factors.

- Higgins, Russell, and Hoffman (2005) used regression to predict raw test scores using a variety of related measures or examinee characteristics and mode of test. The measures used were computer fluidity, computer literacy, home computer use, and school computer use. The examinee characteristics used were gender and whether the student had an individualized education plan (IEP).
- Russell and Plati (2001) used regression to predict constructed-response trait scores and total scores using mode and other variables (typing speed, mid-term English grades).
- Russell (1999) examined typing speed as well as gender.

² This study was not included in detail since it did not provide enough methodological details of the design to warrant inclusion.

- Way and Fitzpatrick (2006) examined the relationship of examinee self-rated computer skills and computer use and student performance using ANCOVA.
- Sandine et al. (2005) conducted a series of ANOVAs on raw scores with a host of other factors (gender, race, parental education level, region of country, school type, type of computer used, typing speed, typing accuracy, editing skill).
- Pomplun and Custer (2005) presented the mean difference of examinees receiving free lunch and those not receiving free lunch by mode. No statistical analyses were conducted on the means.
- Poggio, Glasnapp, Yang, and Poggio (2005) presented the mean difference across a number of dimensions: gender, SES (indicated by examinee receiving no lunch support, free lunch, or reduced lunch), and academic placement category (general education, gifted, or special education).

Summary and Recommendations

The successful evaluation of the comparability of computer-based and paper-based test scores requires a strong inference study design with a focus on key comparability issues. Researchers have used a variety of designs and analyses to examine different aspects of comparability. Table 8 outlines the key issues to consider when designing a study that can generalize to the population of interest and can best identify the impact of testing mode on score and construct equivalence. It is expected that these issues would be addressed in any comparability study. Table 9 presents key comparability questions and hypotheses, organized under the areas of score and construct equivalence. Table 9 also suggests the types of designs and analyses that could be used to test hypotheses.

Because K-12 comparability studies occur in a social and political context, their design and implementation will be shaped by administrative conditions that may conflict with a strong inference design. The investigator must often deal with volunteer samples, cannot control all variables, cannot use random assignment, and/or cannot use a within-subjects design. While these limitations influence the ability to identify the impact of mode and the generalizability of the results, such studies are still important to pursue. The investigator should disclose any design limitations and provide an analysis of the results in light of those limitations.

The assessment of whether scores from a paper-based test are comparable to a computer-based test is, finally, a matter of judgment. The evidence used to support the hypotheses is either judgmental or statistical. In the case of expert review, the investigator needs to interpret the results in light of problems with the limitations of human judgment. In the case of statistical methods, the interpretation involves estimates of both statistical and practical significance. Presumably, the final assessment occurs when a satisfactory amount of evidence that supports the hypotheses has been collected, and little or no evidence contradicts the hypotheses.

Table 8. Issues to Consider in Comparability Study Design

Validity Issue	Suggested Design Features
Sampling	<ul style="list-style-type: none"> • Identify the population from which to sample (e.g., general population, ESL students, IEP students). • Recruit a sample of sufficient size for statistics used. • Outline incentives for school or examinee participation. • Monitor attrition rates. • Present characteristics of participants <i>and</i> non-participants. • Use counterbalancing in within-subjects designs. • Assign examinees to groups to minimize nuisance effects (using random assignment, matching, post-hoc statistical methods). • Monitor the assignment of examinees to groups to ensure sampling plan is properly implemented.
Instrumentation	<ul style="list-style-type: none"> • Study a test that is comparable to one used in practice. • Study a test with enough items to adequately represent the construct. • Ensure that the CBT is built to the same test specifications as the PBT. • Monitor extent to which items are modified for the CBT. • Consider influence of item type (multiple choice, constructed response). • Report on test presentation features (e.g., ability to review items, number of items presented on the screen). • Report on item presentation features (i.e., ability to scroll). • Identify editing features (e.g., grammar- or spell-checking) permitted.
Administration	<ul style="list-style-type: none"> • Ensure test administration conditions are similar to those in practice. • Provide practice tests or a tutorial for CBT examinees. • Provide a tutorial or other training for test administrators. • Ensure computers have the appropriate software and hardware. • Consider examinee access to computers. • Consider bandwidth issues. • Record technical problems of the CBT administration. • Ensure examinee motivation levels are reasonably similar to those in practice. • Administer tests to groups within a reasonably similar time period.
Scoring	<ul style="list-style-type: none"> • Use scoring methods that reflect those used in practice. • Identify the type of score used (e.g., raw, scale, IRT theta). • Outline the equating process for scores and item parameters. • Present inter-rater reliability of constructed responses, if applicable. • Report measurement precision statistics. • Report score distribution information. • Report performance category assignment, if applicable. • Disaggregate data if there is reason to suspect group differences.

Table 9. Study Design and Analyses by Type of Comparability Question

Score comparability		
Focus	Design	Suggested Analyses
<ul style="list-style-type: none"> The scores should have the same measurement precision. 	BS or WS	<ul style="list-style-type: none"> Overall reliability values Overall Standard Error of Measurement Conditional SEM / Test information
<ul style="list-style-type: none"> The score distributions should differ only in difficulty, and hence, be equitable. 	BS or WS	<ul style="list-style-type: none"> Frequency distribution of scores or histogram Measures of central tendency and dispersion (mean, standard deviation) RS-SS tables Distribution of examinees assigned to performance levels
Construct comparability		
Focus	Design	Suggested Analyses
<ul style="list-style-type: none"> The test content and content specifications should be comparable. 	None, BS or WS	<ul style="list-style-type: none"> Blueprint comparisons Expert review of item CBT modifications which may require additional skills Comparisons of different test formats
<ul style="list-style-type: none"> The scores should have the same factor structure. 	BS or WS	<ul style="list-style-type: none"> Tests of dimensionality at item level Tests of dimensionality at item parcel (item groups) level Item level comparisons of CBT and PBT parameters (IRT, CTT)
<ul style="list-style-type: none"> The scores should be highly related to one another. 	WS	<ul style="list-style-type: none"> Correlation (corrected and uncorrected for unreliability) Agreement of assignment into performance levels
<ul style="list-style-type: none"> The scores should have the same relationship to other related measures. 	BS or WS, C	<ul style="list-style-type: none"> Correlation (corrected and uncorrected for reliability) Theory should drive inclusion of measures (e.g., test anxiety influencing PBT scores) Other measures can include tests of ability, attitude measures, examinee characteristics (such as gender and race)

Design: WS=Within-subjects, BS=Between-subjects, C=Criterion

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.
- American Psychological Association (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.
- Bennett, R.E. (2003). *Online Assessment and the Comparability of Score Meaning (ETS Research Report RM-03-05)*. Princeton, NJ: Educational Testing Service.
- Bergstrom, B.A. (1992). *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, CA.
- Boo, J. (1997). *Computerized versus Paper-and-Pencil Assessment of Educational Development: Score Comparability and Examinee Preferences*. Unpublished doctoral dissertation, University of Iowa.
- Campbell, D. T. & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (ed.), *Handbook of Research on Teaching*. Chicago: Rand McNally.
- Choi, S.W., & Tinkler, T. (2002, April). *Evaluating comparability of paper-and-pencil 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.
- Court, S. (2006, April). *The interchangeability of dual-mode testing results (CBT vs. PPT)*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- D'Agostino, R.B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, *17*, 2265-2281.
- Educational Research Center (2006, May 4). Technology Counts 2006: The information Edge. *Education Week*, *25*(35).
- Eignor, D.R. (1993). *Deriving comparable scores for computer adaptive and conventional tests: An example using the SAT*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

- Fitzpatrick, S., & Triscari, R. (2005). *Comparability studies of the Virginia computer-delivered tests*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Gaskill, J. (2006). *Comparisons Between Paper- and Computer-Based Tests: A Literature Review*. Kelowna, BC, Canada: Society for the Advancement of Excellence in Education.
- Higgins, J., Russell, M., & Hoffman, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *The Journal of Technology, Learning, and Assessment*, 3(4).
- Hollenbeck, K., Tindal, G., Steiber, S., & Harniss, M. (1999). *Handwritten vs. word processed statewide compositions: Do judges rate them differently?* Unpublished manuscript, University of Oregon, BRT. Retrieved November 30, 2006 from http://brt.uoregon.edu/files/Hdwrtn_vs_Typed.pdf
- International Test Commission. (2005). *International guidelines on computer-based and internet-delivered test*. Granada, Spain: International Test Commission.
- Johnson, M., and Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *The Journal of Technology, Learning, and Assessment*, 4(5).
- 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.
- Kim, J.-P. (1999). *Meta-analysis of equivalence of computerized and p and p tests on ability measures*. Paper presented at the Annual Meeting of the Mid-Western Educational Research Association, Chicago, IL.
- Mead, A.D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114(3), 449-458.
- Moreno, K.E., & Segall, D.O. (1997). Reliability and construct validity of CAT-ASVAB. In W.A. Sands, B.K. Waters., and J.R. McBride (Eds.), *Computerized Adaptive Testing: From inquiry to operation* (pp.169-174). Washington, DC: American Psychological Association.
- Nuemann, G., & Baydoun, R. (1998) Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement*, 22(1), 71-83.
- Olson, J.B., Maynes, D.D., Slawson, D., & Ho, K. (1989). Comparison of paper-administered, computer-administered and computerized achievement tests. *Journal of Educational Computing Research*, 5, 311-326.

Olson, L. (2003, May 8). Legal twists, digital turns: Computerized testing feels the impact of "No Child Left Behind." *Education Week* 12(35), 11-14, 16.

Paek, P. (2005). *Recent Trends in Comparability Studies (PEM Research Report 05-05)*. Austin, TX: Pearson Educational Measurement.

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. *The Journal of Technology, Learning, and Assessment*, 3(6).

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *The Journal of Technology, Learning, and Assessment*, 2(6).

Pomplun, M., Frey, S., Becker, D., & Hughes, K. (2000). *The validity of a computerized measure of reading rate*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Pomplun, M., & Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K-3 reading tests. *Journal of Educational Computing Research*, 32(2), 153-166.

Rosenbaum, P.R. (1995). *Observational studies*. New York: Springer-Verlag.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Rubin, D.B. (2006). *Matched sampling for causal effects*. New York: Cambridge University Press.

Rudner, L. M., & Peyton, J. (2006, May). *Consider Propensity Scores to Compare Treatments*. Research Report RR-06-07. Graduate Management Admissions Council, McLean, VA.

Russell, M. (1999). Testing on computers: a follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives (online)*. Retrieved November 30, 2006 from <http://epaa.asu.edu/epaa/v7n20/>

Russell, M., & Haney, W. (1997). Testing Writing on Computers: Results of a Pilot Study to Compare Student Writing Test Performance via Computer or Via Paper-and-Pencil. *Educational Policy Analysis Archives*, 5(3).

Russell, M., & Plati, T. (2001). Effects of computer versus paper administration of a state mandated writing assessment. *Teachers College Record*. Retrieved November 30, 2006 from <http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1012andcontext=intasc>

Russell, M. & Tao, W. (2004). Effects of handwriting and computer-print on composition scores: a follow-up to Powers, Fowles, Farnum, & Ramsey. *Practical Assessment, Research and*

Evaluation, 9(1). Retrieved November 30, 2006 from <http://PAREonline.net/getvn.asp?v=9andn=1>.

- Sandine, B., Horkay, N., Bennett, R.E., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online Assessment in Mathematics and Writing: Reports from the NAEP Technology-Based Assessment Project, Research, and Development Series (NCES 2005—457)*. U.S. Department of Education, National Center for Education Statistics. Washington, DC. U.S. Government Printing Office.
- Schwarz, R.D., Rich, C., & Podrabsky, R. (2003). *A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
- Way, W.D., & Fitzpatrick, S. (2006). *Essay responses in 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.
- 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.
- Zhang, L., & Lau, C.A. (2006, April). *A comparison study of testing mode using multiple-choice and constructed-response items – Lessons learned from a pilot study*. Paper presented at the Annual Meeting of the American Educational Association, San Francisco, CA.