A Review of Literature on the Comparability of Scores Obtained from Examinees on Computer-based and Paper-based Tests

Texas Education Agency (TEA) Technical Report Series

November 17, 2008

Executive Summary

- Whenever paper- and computer-based assessments of the same content are administered, professional testing standards and federal accountability both require evidence showing comparability of test scores obtained in the two administration modes.
- Comparability studies have used three main data collection designs: common person, random-equivalent groups, and quasi-experimental. Each design has its advantages and disadvantages (summarized in Table 1), and design selection is typically determined by the specific needs and constraints of each assessment program.
- Comparability studies generally analyze data at the *test level* to evaluate whether the test scores obtained from computer- and paper-based tests are interchangeable. Several studies have also conducted *item-level analyses* to identify item types or characteristics that may perform differently under the two modes of administration.
- Most comparability studies conducted across a number of different state testing programs have found test scores to be comparable across the two administration modes. However, the findings varied in degrees across the different content areas.
- Comparability studies in Texas for the TAKS program have yielded mixed results, but tend to suggest mode effects in mathematics and reading/English language arts.
- Mode effects related to specific item characteristics, such as the need to scroll through a large body of text, have been consistently found for multiple-choice items. Also, students generally perform better on constructed response and essay items when the mode of assessment matches the mode of instruction.
- Based on the findings in this literature review, it is recommended that comparability studies continue to be conducted to help ensure the defensibility of the testing programs in Texas and

in other statewide assessment programs.

Table of Contents

Introduction

The Texas Education Agency (TEA) is pursuing initiatives to include computer-based testing (also known as online testing, electronic testing, or eTesting) into most of its assessment programs. The motivations for moving to computer-based assessments include greater flexibility in administration, reduced administration burdens on district personnel, and the possibility of faster score reporting. In general, the movement toward electronic testing in K–12 assessment programs is picking up momentum as schools increase their technology capabilities and students become more comfortable using the computer for a variety of educational tasks.

However, in most states pursuing computer-based testing (including Texas), many schools do not have the infrastructure and equipment to test all of their students by computer. For this reason, paper and computer-based versions of the same tests are typically offered during the same test administration. Whenever paper and computer-based assessments of the same content are both administered, professional testing standards indicate the need to study *comparability* across paper and computer-based administrations (APA, 1986; AERA, APA, NCME, 1999, Standard 4.10). The No Child Left Behind (NCLB) standards and assessments peer review provided by the U. S. Department of Education (USDE) also requires evidence of comparability as cited in critical element 4.4b: "If the State administers both an online and paper and pencil test, has the State documented the *comparability* of the electronic and paper forms of the test?" As such, a large number of studies, known as *comparability studies*, have been conducted in the past two decades to address this issue.

The purpose of this paper is to provide a review of the literature on comparability studies. The first section gives an overview of the study designs and analysis procedures that are typically employed in comparability studies. The second section summarizes the results that have

been found across comparability studies by content area as well as by item types. The third section describes the history and role of comparability studies in the Texas assessment programs. The paper concludes with recommendations for Texas policymakers as the state moves forward with computer-based testing.

Methods for conducting Comparability Studies

This section focuses on the "how" of conducting comparability studies. It should be acknowledged that the process for conducting comparability studies can be a complicated one and typically requires the involvement of someone with research training and access to specialized statistical analysis software. However, it is important for policymakers to gain an understanding of the key decisions which need to be made in this process and the factors that might influence these decisions. The two major decisions which need to be made are: 1) how to collect the data; and 2) what type of analysis to do with the data once it is collected. To provide a better understanding of comparability methodologies, this section covers how studies commonly address these two major decisions. The first part outlines three typical *data collection designs*. The second part describes typical *types of analyses* utilized for evaluating the comparability of entire tests as well as specific items.

Data Collection Designs

The first, and most critical, step in any comparability study is to collect good data on which to conduct analyses and base conclusions. Without good data, the conclusions that can be drawn about test score comparability will be limited (garbage in, garbage out!). It is more challenging to collect good data than most people think. It is not simply enough to have data. It is not even enough to have large amounts of data. The data collected must have certain features and be collected in very specific ways so that valid comparisons across testing modes can be made.

The following three data collection designs have been implemented in studies to make such comparisons: common person, randomly equivalent groups, and quasi-experimental designs.

Common Person

A *common person* study design is one in which the same people take the test both on computer and on paper. This is also frequently called a *test–retest* study design. Because the same students are taking the test twice, the expectation is that, on average, student scores from the computer-based version would be the same as their scores from the paper version. Test–retest studies are generally conducted using a counter-balanced order, meaning that some students take the computer test first while others take the paper test first. Counter-balancing helps to mitigate effects of practice and order on test performance.

This type of study design was used by the National Association of Educational Progress (NAEP) to assess the comparability of writing in 2002 (Horkay, Bennett, Allen, Kaplan, & Yan, 2006). A subset of students who completed the main NAEP administration on paper was selected to test again on computer. The states of Kansas (Poggio, Glasnapp, Yang, & Poggio, 2005) and Oregon (Oregon Department of Education, 2007) have also conducted comparability studies using a common person design. The state of Kansas conducted its comparability study on mathematics for grade 7 students; while the Oregon Department of Education explored comparability of test scores in mathematics, reading and science.

The advantages of a common person design are that a smaller sample of students is needed and that it is a very powerful design for detecting differences. Because the same students test twice in a common person design, fewer students need to be included in the study overall. In essence, the sample size requirements can be cut in half because each group (paper and computer) consists of the same students. In addition, the common person study design is more

likely to detect small differences between test modes because the influence of additional factors (e.g. gender, ability, SES, etc.) is eliminated. The groups taking the test in each mode are literally identical.

The disadvantages of a common person design are that students are required to test twice and that factors associated with testing twice must be considered. Asking students to test twice creates additional test burden to the student. Concerns about motivation, fatigue, and boredom must be considered. One strategy to increase motivation has been to have students test in both modes before receiving scores from either mode and allowing students to use the higher of the two test scores. Other challenges involve practice effects if the same test is used twice, the creation of truly equivalent forms if different forms are to be used, and the effect of learning between the time the students take the first and second tests.

Randomly Equivalent Groups

When it is not feasible to test students multiple times, another study design called *randomly equivalent groups* can be utilized. This study design includes the use of two groups of students who are assumed to be the same on all characteristics that could be related to test performance (subject matter proficiency, grade level, educational experience, demographic characteristics, etc.). One group takes the test on computer, and the other group takes the test on paper. One way to create “equivalent” groups is to randomly assign students to either the computer group or the paper group. Because of random assignment, there should be no systematic differences between the groups. As such, the sample of students in each group is expected to demonstrate equal performance on the test. Any differences in performance between the computer and paper groups could therefore be attributed to the mode of administration.

The 2001 NAEP administration in mathematics (Bennett, Braswell, Oranje, et al., 2008) used a randomly equivalent groups design to evaluate comparability between the computer-based and paper versions of those tests. NAEP randomly assigned students within the same school to test either by computer or by paper.

The main advantages of a study designed using random assignment are that students need to test only once, no additional manipulation or evaluation of the groups is required, and calculations of performance differences can be quite simple. Because students are assigned to test in one mode or the other, students need to complete the test only once. This can overcome some of the concerns about the common person design including the effects of fatigue and practice. Additionally, due to the fact that the two groups of students are assumed to be the same on all important characteristics, no additional manipulation or evaluation of the groups is required. Straightforward calculations (like the difference in average scores between the two groups) can be used to determine any performance differences.

The primary disadvantage of random assignment is that it is difficult to implement. In order to randomly assign students to conditions, a list of all students needs to be available. Furthermore, schools and students have to be willing to test each student in the condition to which he or she was assigned. In a high-stakes testing situation, schools may be reluctant to test students in a non-preferred mode.

Quasi-experimental Designs

Although strictly controlled experiments to study comparability using a counterbalanced test–retest design or random assignment seem like a good idea, carrying out these types of controlled experiments within the context of a statewide assessment program is difficult. For this reason, quasi-experimental designs are another viable study design option. A quasi-experimental

design is one in which there are two groups (e.g., computer and paper), but students have not been randomly assigned to a group. Therefore, the groups may not be equivalent at the time at which data are collected. During data analysis, steps are taken to account for any pre-existing differences between the two groups so that the single variable of interest (i.e., mode of administration) can be studied. In other words, statistical techniques are used to create equivalent groups from non-equivalent groups. Some of the specific types of data analysis procedures that are used with a non-equivalent groups design are presented in the section on types of analyses.

Since 2002, students have been able to choose whether they wanted to take the Praxis™ Pre-Professional Skills Test by paper-and-pencil or on computer (Yu, Livingston, Larkin, & Bonett, 2004). Similarly, in the state of Texas, participation in the computer-based version of the Texas Assessment of Knowledge and Skills (TAKS) is voluntary (Way, Davis, & Fitzpatrick, 2006). As such, comparability studies for both of these assessments were cases of quasi-experimental designs, and special data analyses procedures were used.

The advantages of quasi-experimental designs are that students need to test only once and that the mode of testing can be selected by the school or student. A quasi-experimental design is one that is fairly straightforward to implement in a statewide testing program because students need to test only once, which could easily be part of the regular testing administration. Additionally, schools or students can select in which mode to administer the test depending on the skills and interests of the students and the technology resources available at the school.

The primary disadvantage of quasi-experimental designs is that it relies on an ability to create equivalent groups from groups that initially may not be the same. This design requires that additional information about each student be available for use in creating equivalent groups. This is usually done by matching the students from each group on some important variables that are

highly related to the test scores of interest (e.g., gender, ability, SES, etc.). Specialized statistical techniques are used to create similar groups, and the quality of the study results is dependent on the data available and techniques used to create the similar groups.

The main advantages and disadvantages of the three data collection designs are summarized in Table 1 below.

Table 1: Advantages and Disadvantages of Data Collection Designs

Data Collection Design	Advantages	Disadvantages
Common Person	• Smaller sample size required • Powerful design for detecting differences because influence of student characteristics is controlled	• Students have to test twice • Effects of motivation, fatigue, boredom, practice, time, and learning must be considered
Equivalent Groups	• Students test only once • No additional group manipulation or evaluation • Simple calculations for analysis of data	• Random assignment of students presents practical challenges
Quasi-experimental	• Students test only once • Mode of testing can be selected by the school or student	• Requires additional student information and statistical techniques in an attempt to create equivalent groups • Quality of the results are dependent on the ability to create equivalent groups

Types of Analyses

Different data collection designs can lead to different methods of data analysis. This review of analysis methods is divided into two parts to reflect two different possible levels of comparability analyses. The first part addresses *test-level* comparability, which considers whether or not student scores on the total test are comparable between computer and paper. The second part focuses on *item-level* comparability. This type of analysis breaks the level of study down to individual items to determine if certain types of items or certain qualities about items lead to differences in performance between computer and paper.

Test-Level Analyses

Analyses at the total test-score level are usually the first, and sometimes only, step in comparability research. These types of analyses are implemented to answer the bottom-line question: are student scores from the computer-based test comparable to student scores from the paper-based test? While there are many ways to assess comparability (see Paek, 2005), two common types of test-level analyses will be described in detail: mean differences and item response theory (IRT) analyses.

Mean differences. One of the most common and straightforward ways to evaluate comparability is to compare the mean test scores between the two modes. When either a common person or randomly-equivalent groups data collection design is used, it can be expected that the average test score in the computer mode would be similar to the average test score in the paper mode. By directly comparing the average test scores of the two groups, it can be determined if test performance differs significantly based on the mode of administration**.**

The mean difference method is straightforward, easy to calculate, and easy to explain. However, this method assumes that there is a constant mode effect for all students, which may or may not be reasonable. For example, it might be that lower ability students show a greater difference between online and paper performance than higher ability students. However, in calculating only a mean difference, the difference in performance due to test mode would be assumed to be the same for the low- and high-ability students. The effect would likely be underestimated for lower ability students and overestimated for higher-ability students.

In a study of a southeastern state's end-of-course algebra and biology assessments, mean differences in scale scores were analyzed as one indicator of comparability in a common person study design (Kim & Huynh, 2007). Results showed that students taking the algebra test on

paper scored higher than those testing on computer. There were no differences in test performance based on mode of administration for biology.

IRT-based analyses. IRT models are specialized statistical models used in test development to estimate information about test items (e.g., item difficulty) as well as information about test-takers (e.g., student ability). In the most basic IRT model (the one-parameter logistic, or Rasch model), test items are assumed to differ from each other only in item difficulty. It is assumed that students are more likely to answer an easier item correctly than a difficult one, regardless of the ability level of the student. Likewise, a student with higher ability has a greater probability of correctly answering an item than a student of lower ability, regardless of the difficulty of the item. In more complex IRT models, additional variables can be added (e.g., guessing, etc.). There are a variety of ways that IRT models can be used to evaluate comparability.

IRT-based analyses are commonly used to compare different test forms in the same testing mode given over time. Similar analyses can be conducted to compare test forms given in different modes if a common person or randomly equivalent groups data collection design was implemented. Large sample sizes are required in order to accurately estimate item difficulty, student ability, etc. The method is more statistically sophisticated than the mean difference method, and may be more difficult to explain to the general public.

The state of Oregon used an IRT-based method along with a common person study design to evaluate comparability of their Technology Enhanced Student Assessment (TESA; Oregon Department of Education, 2007). Because the same students took the computer and paper versions of the Oregon state assessment, their IRT ability estimates were set to be the same between both modes. Using those abilities, the item difficulties and total test scores could be

estimated separately for items on the computer and paper versions of the test. Differences in total test scores between modes would indicate a test-level mode effect. For the Oregon study, no statistically significant test-level differences were found across the two administration modes.

The NAEP mathematics study also used an IRT-based approach but came at the problem from the opposite angle (Bennett et al., 2008). In this study, the estimated information about the items was set to be the same for the computer-based and paper versions. Then student ability was estimated separately for the computer version and paper version of the test using the same item estimates for both. Differences in the estimated abilities between the two groups would indicate a mode effect.

Note that the two types of test-level analyses described above can be used when the data are collected under the common person or equivalent groups design. When the data collection design is quasi-experimental, special types of analysis are needed to manipulate or transform the data collected. Two such methods found in comparability literature include the matched samples comparability analysis and propensity scoring.

Matched Samples Comparability Analysis. The matched samples comparability analysis (MSCA; Way, Um, Lin, & McClarty, 2007; Way et al., 2006) does not require random assignment to test administration mode or require students to test twice. As such, it has been a popular methodology for several states where students or schools self-select in which mode to test. The method assumes that all students taking the test by computer or on paper have available previous or concurrent test scores in a content area related to the content area under study, as well as additional demographic information (e.g., gender, ethnicity). The method proceeds by drawing a sample of students testing by computer and matching them to students taking the paper form (or vice versa) so that the two samples have identical profiles of previous test scores

and demographic variables. Computer and paper performance can then be compared between the two matched groups, and the sampling and comparisons are repeated for some number of times (e.g., 100 or 500).

Like the IRT-based methods, the MSCA requires a large sample of students testing so that appropriate matches can be found, and the analyses are not transparent. In addition, matching variables (e.g. previous test scores and demographic information) must be available for each student as well. The quality of the results depends on the quality of the available characteristics on which students can be matched. Despite these challenges, the MSCA approach can be used with quasi-experimental designs, so schools and student can select in which mode to test. Texas, Michigan, South Carolina, Arizona, and Maryland have all conducted comparability studies using the MSCA method. The comparability results have been mixed across states and across test administrations (Way, Lin, & Kong, 2008).

Propensity Scoring. Propensity scoring is another method that can be used for a quasi-experimental study design (Rosenbaum & Rubin, 1985). In order to calculate a propensity score, available demographic variables are used to predict whether the student tested on computer or on paper. The propensity score is the weighted sum of the predictor variables. Each student is assigned a propensity score by assigning the greatest weight to the variables on which the groups (computer vs. paper) are most different and the smallest weight to the variables on which the groups are most similar. This information is used to create similar groups between online and paper (based on demographics) and the mean scores can be compared between the similar groups.

This method has much in common with the MSCA. Both methods create equivalent groups and therefore can be used with quasi-experimental designs. Both require additional

student-level variables. Although the propensity scoring method does not require previous test scores, it does require more demographic variables than the MSCA. Also, because the ultimate evaluation is that of mean differences between the two groups, a constant mode effect is assumed. As was discussed earlier for the mean differences method, it may not be reasonable to assume that all students show the same performance difference between online and paper. Yu et al. (2004) used propensity scoring to investigate differences between computer-based and handwritten essays and used variables from eight different categories as predictors in their analysis.

A summary of some of the features of the different test-level analysis methods is presented in Table 2.

Table 2: Summary of features for the different test-level analyses

Features	Mean Differences	IRT-based Approaches	MSCA	Propensity Scoring
Use with common persons designs	X	X		
Use with randomly equivalent groups designs	X	X		
Use with quasi-experimental designs			X	X
Transparent, easy to explain method	X			
Large sample size required		X	X	
Assumes constant mode effect	X			X
Previous test scores needed			X	
Student demographic information needed			X	X
Allows for score adjustments	X	X	X	X

Item-Level Analyses

Test-level analyses, however, only tell part of the story. When data collection designs permit giving the same items to the groups testing on computer and by paper-and-pencil, item-level analyses are useful to determine if there are mode effects for individual items. (In test–

retest data collection designs, item-level analyses are usually not possible.) This can be a first step in understanding why mode differences may occur and how changes to the items or the presentation of items may minimize or even eliminate mode differences. Further investigation of items that do show differences may help test developers understand the cause of the mode effects. For example, in a study of reading passage interfaces, Pommerich (2004) hypothesized that some of the performance differences for certain items may be due to highlighting of text or the layout of the passages. Item-level analyses can be conducted before or after the test has been administered, or both. This section will provide information about five methods of item-level analysis: content analysis, mean differences, response distributions, IRT-based differences, and differential item functioning.

Content analysis. Prior to conducting a comparability field study, NAEP performed a content review of the items in the bank (Sandene, Horkay, Bennett, et al., 2005). The test developers and content staff felt that the majority of items could be presented on the computer in a manner similar to the paper version of the test with little or no problems. There were a few items, however, which the content reviewers felt were difficult to transfer to computerized administration and were less likely to be appropriate for computer-based administration.

Mean differences. Similar to the test-level analysis, one of the most common ways to look for item-level mode effects is to evaluate differences in the percentage of students who answered the item correctly in each mode. If the percentage of students answering the item correctly on paper is similar to the percentage of students answering the item correctly on computer, then the item does not show any mode effects.

IRT-based differences. A different way of estimating item difficulty (other than percentage of students answering correctly) is through the use of IRT. As described earlier, IRT

models are used to estimate information about test items as well as information about test-takers. When applying IRT to item-level comparability analyses, researchers can compare differences between the difficulty estimates for items presented on computer and presented on paper. Using this technique, Bennett et al. (2008), showed that although the rank order of item difficulties in each mode was similar, the NAEP math items presented on computer all tended to be slightly more difficult than the same items presented on paper.

Differential item functioning. Differential item functioning (DIF) is a statistical term used to indicate that two groups perform differently on an item (after controlling for differences in abilities of the groups). In the case of comparability research, DIF analysis is used to detect differences in item performance between the computer and paper administration modes. Research conducted for the states of Kansas (Poggio et al., 2005) and Texas (Keng, McClarty, & Davis, 2008) have both conducted DIF analysis as one part of their comparability studies. The state of Kansas found a few mathematics items (9 out of 204 items) that behave differently across the two modes; while the Texas study found mode effects for several English language arts (ELA) and mathematics items.

Response distributions. Another source of information for assessing item-level mode differences is to look at the percentage of students choosing each response option (response distributions) for test items between the two modes. In comparing response distributions, differences are calculated not only for the percentage of students selecting the correct answer but also for the percentage of students selecting each of the incorrect answers. A response distribution analysis looks to see if the pattern of responses differs between the computer and paper administration modes. In a response distribution analysis of grade 11 mathematics items, Keng et al. (2008) noted that the item layout on screen may have affected the student response

patterns. One item on which the students testing by computer performed better contained several large graphics, so that only the first two answer choices could be seen on the initial on-screen presentation of the item. To see the last two answer choices, the student would need to scroll down through the item. In the paper version of the item, all four answer choices could be seen at the same time. Because the correct answer was one of the first two answer choices, the researchers hypothesized that the students testing on computer may have ignored the answer choices that could not be seen and selected their answer from the two viewable alternatives (instead of the four viewable alternatives on paper).

Decision Making: What do you do with the results of comparability studies?

This is a critical element in conducting and evaluating comparability studies. Once the data collection design has been determined and data has been collected and analyzed, what happens next? If the results show that the test performance across the two testing modes is comparable, the conclusion is that the scores on the two modes can be used interchangeably. Students are not advantaged or disadvantaged by the mode in which the test is administered.

The more difficult question, however, is what to do if the results of the study show that the tests are not comparable. Whether or not it makes sense to adjust test scores to make them more comparable depends on the data collection design. For example, with a test–retest design, score adjustments are not necessary. With random groups or quasi-experimental designs, score adjustments may be possible if the comparability study can be completed before scores are reported. In this case, policymakers can decide whether or not to make an adjustment to the test scores. If a score adjustment is to be made, there should be defensible evidence in support of such a change. Under these circumstances, the choice of test-level analysis method is extremely

important. Each of the different test-level analyses described earlier provides different kinds of information to use if the test modes aren't comparable.

- An analysis of *mean differences* provides a single value that is the average difference between the two groups. For example, if the mean on paper was 67.5 and the mean on computer was 68.5, the mean difference would be 1.0 in favor of the computer. In order to adjust test scores to make them more comparable, one point could be added to every student that tested on paper. This adjustment would be made across the board for every student, as there is no way of knowing from the mean differences whether the mode effect really impacted some students more than others.
- An *IRT-based analysis* results in either two sets of item statistics (one from paper and one from computer) or two sets of ability estimates (one from paper and one from computer). In either case, these values can be used to compute a new score for the student. This method does not require a constant adjustment, but rather the adjustment can be tailored based on the items and/or students that were impacted greatest by the mode of administration. Two sets of scoring tables can be developed, one for paper and one for computer.
- The results from *MSCA* also can provide two sets of scoring tables. The results of the analysis are provided in the same way that scores are reported. For example, there may be one set of raw score to scale score conversion tables for paper. The results of the MSCA analysis will provide an alternate raw score to scale score conversion table for computer-based testers. This alternate table is able to take into account whether low ability students were influenced by the mode of testing more than high ability students.
- A *propensity score* analysis ultimately results in the same available statistics as the mean differences approach. A mean difference can be calculated from the two similar groups, and

this mean difference could be applied to everyone in, for example, the paper group to make their scores more comparable to the computer group.

Results of Previous Comparability Studies

This section focuses on "what" has been found in comparability studies. As policymakers consider the implications that testing in two modes will have on test score comparability in their state, it may be helpful to know the outcomes of comparability research in other states. As one might suspect, the outcomes of comparability research may vary by content area (reading, mathematics, science, etc.), item type (multiple choice vs. constructed response), and testing program (different states assessment programs, licensure and credential tests etc). Results of comparability research are not always consistent even across tests in the same subject area and with similar features. This makes it difficult to predict the result of a comparability study for a specific state's assessment. However, there are some general trends which can be observed across the extensive body of research in this area.

This section summarizes the results of comparability studies that have been conducted in the past twenty years. The results are presented first by content areas and then by item types. In addition, hypotheses for why computer-based and paper tests are not comparable, as suggested by the various studies, are provided.

Results by content area

Mathematics

Table 3 below summarizes the study findings for comparability studies that examined tests in mathematics content area. The studies included are those which compared student

performance in the computer and paper administration modes at the *test level*; that is, whether the students' overall test scores were comparable across the two modes.

The studies reviewed included tests from various grade levels from elementary to high school. A few studies also looked at specific mathematics domains in high school (end-of-course), such as algebra and geometry. Of the twenty-three studies listed, fourteen found that student performance was comparable across the two modes, eight found that the computer-based test was more difficult, and only one study found that the paper test was more difficult. Thus, although the overall findings were not conclusive, the majority of the studies investigated found mathematics tests to be comparable in overall difficulty across the two administration modes.

Table 3: Test-Level Comparability Study Findings for Mathematics

Harder Mode	Studies
Computer (8)	Applegate (1993), Kindergarten Choi & Tinkler (2002), Grade 3 Cerillo & Davis (2004), Algebra Ito & Sykes (2004), Grades 4–12 Russell (1999), Grade 8 Sandene, Horkay, Bennett, Allen, Kaplan & Oranje (2005) Way, Davis, & Fitzpatrick (2006), Grade 11 Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan (2008), Grade 8
Paper (1)	Choi & Tinkler (2002), Grade 10
Comparable (14)	Russell & Haney (1997), Grade 6–8 Kingston (2002), Grades 1, 4, 6, 8 Pearson (2002), Algebra Pearson (2003), Algebra II Wang (2004), Grades 2-5, 7–12 Nichols & Kirkpatrick (2005) Poggio, Glassnapp, Yang, & Poggio (2005), Grade 7 Fitzpatrick & Triscari (2005), Algebra I, Algebra II, Geometry Johnson & Green (2006), 11 year-olds Way, Davis, & Fitzpatrick (2006), Grade 8 Kim & Huynh (2007), Algebra Oregon Department of Education (2007), Grades 3–10 Puhan, Boughton, & Kim (2007) Wang, Jiao, Young, Brooks, & Olson (2007), Grades 2–12 (meta-analysis)

In addition to test-level comparisons, several studies investigate whether there are differences in student performance at the item level for mathematics (Russell & Haney, 1997; Ito & Sykes, 2004; Pommerich, 2004; Poggio et al., 2005; Fitzpatrick & Triscari, 2005; Johnson & Green, 2006; Kim & Huynh, 2007; Puhan et al., 2007; Bennett et al., 2008; Keng et al., 2008). Some studies found no mode differences at the item level (e.g. Fitzpatrick & Triscari, 2005; Kim & Huynh, 2007; Puhan et al., 2007). Others found mode effects, but were unable to identify item characteristics that may have led to the differences (e.g. Poggio et al., 2005).

A few studies found mode effects and were able to identify characteristics of mathematics items that may have led to mode differences (Ito & Sykes, 2004; Johnson & Green, 2006; Bennett et al., 2008; Keng et al., 2008). In addition, Sandene et al. (2005) listed several types of mathematics items that were deemed inappropriate for the computer-based NAEP math assessment. Although the item characteristics in Sandene et al. were not based on formal analyses but on the experience of NAEP staff members, they do correspond to item features identified in other studies. Some characteristics of mathematics items that are hypothesized to lead to mode effects include items that:

- are multipart or require much scrolling to see the entire item
- are intended in part to determine how effectively the student can manipulate some physical tool (such as a ruler or protractor)
- require the student to create drawings, enter a lengthy amount of text, or produce mathematical formulas
- require graphing or geometric manipulations
- require extended tutorials or lengthy item-specific directions for responding

- require paper stimulus materials
- require significant changes to the formatting to render on a computer
- assume a screen resolution that is the same across all student computers

English Language Arts (Reading and Writing)

Table 4 summarizes test-level comparability results for studies that examined tests in the English language arts (ELA), reading, or writing content areas. As with mathematics, the studies involved ELA, reading, or writing tests from early elementary to the high school level. Of the twenty-six studies listed, six of them contain findings that show the computer-based test to be more difficult, four found the paper test to be harder, and the majority (sixteen) of studies found the tests to be comparable across the two administration modes.

Table 4: Test-Level Comparability Study Findings for English Language Arts

Harder Mode	Studies
Computer (6)	Choi & Tinkler (2002), Grade 3 Cerillo & Davis (2004), High School Way, Davis, & Fitzpatrick (2006), Grade 8 Ito & Sykes (2004), Grades 4–12 Yu, Livingston, Larkin, & Bonett (2004), Writing Fitzpatrick & Triscari (2005), High School
Paper (4)	Russell & Haney (1997), Grades 6–8, Writing Choi & Tinkler (2002), Grade 10 Pomplun, Frey, & Becker (2002), High School O'Malley, Kirkpatrick, Sherwood, Burdick, Hsieh & Sanford (2005), Grades 2–5, 8
Comparable (16)	Russell & Haney (1997), Grades 6-8, ELA Russell (1999), Grade 8, ELA Kingston (2002), Grades 1, 4, 6, 8 Pearson (2002), High School Pommerich (2004), Grades 11–12 Wang (2004), Grades 2-5, 7–12 Higgins, Russell, & Hoffman (2005), Grade 4 Michigan Department of Education (2006), Grade 6 Nichols & Kirkpatrick (2005) Horkay, Bennett, Allen, Kaplan, & Yan (2006), Grade 8 Writing

	Kim & Huynh (2006), High School Way, Davis & Fitzpatrick (2006), Grade 11 Oregon Department of Education (2007), Grades 3–10 Pommerich (2007), Grades 11–12 Puhan, Boughton & Kim (2007) Wang, Jiao, Young, Brooks & Olson (20078), Grades 2–12 (meta-analysis)

Of the studies that performed item-level comparability analyses (Russell & Haney, 1997; Ito & Sykes, 2004; Pommerich, 2004; Yu et al., 2004; Higgins et al., 2005; Michigan Department of Education, 2006; Horkay et al., 2006; Pommerich, 2007; Keng et al., 2008), some found either no differences or were unable to identify items characteristics that led to mode differences. However, some research studies (Pommerich, 2004; Yu et al., 2004; Higgins et al., 2005; Keng et al., 2008; Way, Davis, & Strain-Seymour, 2008) did suggest item characteristics of ELA, reading, or writing items that may function differently across the two modes. These findings include:

- Items that required on-screen scrolling or paging because of lengthy passages or item content tended to be more difficult for students testing on computer.
- Subtle features such as the layout of passages, location of line breaks, ease of navigation through items, alignment of item with reading passage, and highlighting of relevant text (by underlining, numbering or with color) seemed to contribute to item-level performance differences.
- Providing computer-based tools, such as highlighters and review markers, tended to help the performance of computer-based test-takers.
- Students may perform better on essay items when the mode of assessment matches the mode of instruction.

Science

Table 5 summarizes the test-level comparability results for studies that investigated one of the science content areas. These studies compared tests from elementary to high school and include some that examine specific domains in science at the high school (end-of-course) level such as biology, chemistry, physical sciences, and earth sciences. Of the fourteen studies listed, ten found the tests to be comparable overall, three found the science tests to be harder in the paper mode, and one found the computer-based version of the science test to be more difficult. Thus, in general, science tests were found to be comparable across the two administration modes.

Table 5: Test-Level Comparability Study Findings for Science

Harder Mode	Studies
Computer (1)	Cerillo & Davis (2004), Biology
Paper (3)	Chin, Donn, & Conry (1991), Grade 10 Russell (1999), Grade 8 Pommerich (2007), Grades 11–12
Comparable (10)	Russell & Haney (1997), Grades 6–8 Kingston (2002), Grades 4, 6, 8 Pearson (2002), Earth Science Pearson (2003), Biology Pommerich (2004), Grades 11–12 Fitzpatrick & Triscari (2005), Biology, Chemistry Kim & Huynh (2006; 2007), Physical Science, Biology Way, Davis, & Fitzpatrick (2006), Grade 11 Oregon Department of Education (2007), Grade 3–10 Pommerich (2007), Grades 11–12

Six of the studies reviewed (Russell & Haney, 1997; Pommerich, 2004; Fitzpatrick & Triscari, 2005; Kim & Huynh, 2007; Pommerich, 2007; Keng et al, 2008) also performed item-level analyses of science items. Only two of the studies (Pommerich 2004; 2007), however, identified significant item-level mode differences and suggested possible item characteristics that led to these differences. Pommerich (2004) attributed the higher student performance on certain

computer-based graphical science items to a *focus effect*. That is, allowing students testing on computer to view only graphics relevant to a specific item on screen and not having the visible distraction of extraneous information from adjacent items helped them perform better on such items. At the same time, Pommerich (2004; 2007) suggested that science items requiring scrolling on screen tended to favor student taking the test on paper, a similar finding in the item-level comparability analyses for mathematics and ELA.

Social Studies

Table 6 gives the summary of test-level comparability studies that examined tests in one of the social studies content domains. Relatively fewer studies have investigated mode differences in social studies compared with the other three content areas. However, like the other content areas, the majority of studies (3 out of 5) reviewed involving social studies found the overall performance of students to be comparable across the two administration modes.

Table 6: Test-Level Comparability Study Findings for Social Studies

Harder Mode	Studies
Computer (1)	Fitzpatrick & Triscari (2005), World Geography
Paper (1)	Fitzpatrick & Triscari (2005), History
Comparable (3)	Kingston (2002), Grades 4, 6, 8 Michigan Department of Education (2006), Grade 6 Way, Davis, & Fitzpatrick (2006), Grades 8, 11

Two studies (Fitzpatrick & Triscari, 2005; Keng, et al., 2008) performed item-level comparability analysis for social studies items. Although items were found with significant mode differences in both studies, in neither study were researchers able to come up with observable characteristics of the social studies items that may have led to the mode differences. Further research into the comparability of social studies items and tests is certainly warranted.

Results by item types

Multiple-Choice Items

Most comparability studies listed in Tables 3-6 examined tests consisting mainly of multiple-choice (MC) items. Several of the content area-specific characteristics of MC items thought to lead to differential item performance across administration modes have already been summarized in the previous section.

One characteristic that has been consistently shown to lead to mode differences in MC items across all content areas is when items are associated with a body of text that requires *scrolling*. Numerous studies have found that if an item could be presented in its entirety on a computer screen, then little or no mode differences existed at the item or test level (Spray, Ackerman, Reckase, & Carlson, 1989; Bergstrom, 1992; Hetter, Segall, & Bloxom, 1997; Bridgeman, Lennon, & Jackenthal, 2001; Choi & Tinkler, 2002). However, when text or passage associated with an item did not fit complete on one screen, significant mode effects were found at the item and overall test levels, favoring the students taking the paper and pencil test (Bergstrom, 1992; Bridgeman et al, 2001; Choi & Tinkler, 2002; Pommerich, 2004; Higgins et al., 2005; O'Malley et al., 2005; Way et al., 2006; Keng et al., 2008).

One study (Pommerich, 2004) also identified *speededness* as a test characteristic that could lead to mode effects in MC items. A test is defined as speeded when students are required to complete their tests within a pre-determined timeframe. These studies found that the difference in performance across modes tended to be greater for items at the end of the test (where student were often running short on time) than at the beginning of the test. Pommerich suggested the "no-bubble effect" as an explanation for this mode difference. When responding to MC items, students testing on computer do not have to bubble in their responses; they only need

to click on their answer choice. Thus, when examinees are pressed for time, this minor difference in time required to rapidly respond to items could lead to mode differences near the end of a speeded test in favor of students testing on computer.

Constructed Response Items

Constructed response (CR) items require the student to create their own responses. Such items may require a variety of student-produced responses, including short written text, numeric responses and geometric drawings. Four of the reviewed studies (Russell, 1999; Johnson & Green, 2006; Michigan Department of Education, 2006; Bennett et al., 2008) compared student performance on CR items. These studies identified two factors that appear to lead to differential performance on CR items across modes. One was test-taker's *degree of familiarity* with typing text responses on the computer. All studies found that students with computer experience, whether through instruction or tests previously taken, tended to perform as well, if not better on the computer-based version of the CR items. Another factor was the feasibility or amount of work required to convert the paper and pencil version of a CR item to the computer version. Results from Bennett et al. suggested that it may sometimes be harder to maintain the level of item difficulty for CR items than for MC items when moving them from paper to computer.

Essay Items

Several studies have investigated differences in performance on essay items administered across the two modes (Sweedler-Brown, 1991; Powers, Fowles, Farnum, & Ramsey, 1992; Russell & Haney, 1997; Bridgeman & Cooper, 1998; Yu et al., 2004; Breland, Lee, & Muraki, 2005; Horkey et al., 2006; Way & Fitzpatrick, 2005; Puhan et al., 2007; Davis, Strain-Seymour, Lin, & Kong, 2008). The two factors found to lead to mode differences for CR items – students' degree of familiarity with taking computer-based tests and the feasibility of converting paper-

based items to computer – were also cited as factors leading to differential mode performances in essay items.

Two additional factors, the difference in *perceptions of essay scorers* between handwritten and computer-typed essay responses and the *computer writing interface*, were also suggested. In some studies, essay scorers frequently rated the same essay more stringently when it was in typed format than when it was handwritten. One suggested explanation for this difference in scorer perception was that scorers might have higher expectations for essays that were typed because these were likely to be perceived as final drafts (Breland et al., 2005). Another explanation was that scorers might unconsciously identify the individuality in handwritten essays more easily than in typed essays (Yu et al., 2004). Other research has shown that typed essay performance was better than paper essay performance when the computer-based testing interface included standard word processing supports (e.g., cut, copy, paste, undo) and a way for students to monitor how much the student had written compared to the amount of space they had (Davis et al., 2008). This interface more closely resembled the word processing experience that students used in the classroom.

Comparability and the Texas Assessment Program

In recent years, several computer-based testing initiatives have been implemented in the Texas assessment programs— the Texas Assessment of Knowledge and Skills (TAKS), the Texas English Language Proficiency Assessment System (TELPAS), and the End-of-Course (EOC) assessments. This section talks specifically about the history and role of comparability studies in the TAKS, TELPAS reading, and EOC assessments.

TAKS

Computer-based testing for TAKS began in 2004 with a pilot study of TAKS grade 8, including reading, mathematics, and social studies. Since that time, TAKS has had computer-based administrations offered in grades 7, 8, 9, 10, and exit level. Currently the computer-based offerings for TAKS include the exit level retests in ELA, mathematics, science, and social studies.

Each time a TAKS test form is offered in both computer-based and paper modes for the first time, a comparability study is conducted after the test administration. To allow districts and campuses the flexibility of choosing their mode of test administration, a quasi-experimental study design has typically been used in TAKS comparability studies. A matched samples comparability analysis (MSCA) is used to evaluate test-level performance differences between the two modes; while mean difference, DIF and response distribution analyses are used to identify items that exhibited significant mode effects. Decisions about adjustments to test scores are made based on the test-level comparability findings. *If no test-level differences are found*, the test scores are considered comparable across testing modes, and no statistical adjustment is needed. *If test-level differences are found*, the scale scores associated with each raw score are adjusted to account for these differences. In practice, this may result in a difference between the raw score a student would need to achieve "Met the Standard" or "Commended Performance" on the computer-based test and the raw score needed on the paper test. This adjustment, if needed, is made prior to reporting scores back to students.

The comparability results from each year's test administrations can be found on TEA's website in the technical digest for that year. A general summary of the results from 2005 to 2008 can be found in Table 7 below. Historically, mode effects have been seen fairly consistently in

reading, ELA, and mathematics. Comparability results for science and social studies have been less consistent.

Table 7: Summary of Computer-based vs. Paper TAKS Comparability Studies from 2005-2008

Subject	No Differences	Differences
Reading	1	5
English language arts	2	5
Mathematics	1	13
Science	4	6
Social Studies	4	6

TELPAS Reading

A computer-based pilot administration of the TELPAS reading field test for grade 2 students was administered during the spring 2007 testing window as a first step in transitioning to a fully computer-based TELPAS reading assessment. In spring 2008, TELPAS reading was administered as both a paper and computer-based assessment, with campuses self-selecting their testing mode. As with TAKS, a quasi-experimental study design was used for the TELPAS comparability study employing MSCA to identify mode effects. The results will be available on the TEA website in fall of 2008. Because TELPAS reading will be a computer-based only assessment beginning in 2009, no further comparability studies will be needed. If the TELPAS program were to begin to offer paper tests in the future (along with a computer-based version), comparability would need to be addressed again.

EOC

In fall 2002, the EOC Algebra I assessment was made available on the computer, and districts were given the option of administering this test in either computer-based or paper format. In spring 2005, a new Algebra I EOC assessment was developed exclusively in computer-based form. In spring 2008, EOC assessments were offered on computer only in

Algebra I, biology, and geometry as live tests with score reports available 24 hours after testing and in chemistry and US History as field tests. As EOC moves toward being the assessment program required for high school graduation in Texas, not all schools may have the technology infrastructure required to test all of its students on computer. In preparation, the World Geography field test will be administered in both paper and computer-based formats to campuses selected for the field test in spring 2009. A comparability study will be conducted following the field-test administration.

Conclusions

What can we surmise from the large body of literature on the comparability of computer-based and paper assessments? Because the majority of comparability studies in other states have found the computer- and paper-based versions of their tests to be comparable overall (see Tables 3–6), a natural question to ask is: have we amassed enough evidence (or will be ever get to such a point in the near future) to say that computer- and paper-based tests are comparable so that no more comparability studies are necessary? The answer depends on the specific needs and circumstances of each testing program. Each state needs to assess its situation and weigh the costs of conducting regular comparability studies against the risks of not conducting them.

In Texas, when computer- and paper-based versions of the same test are available for a particular operational administration, districts and campuses can choose the mode in which they would like their tests administered. Because of this, comparability studies in Texas have typically been quasi-experimental designs. Both test- and item-level analysis are usually conducted to detect mode effects in test scores as well as to identify item characteristics that perform differently across the two modes. Mode effects have been found fairly consistently on the TAKS assessments in mathematics and reading/English language arts.

Comparability studies have generally been conducted each time the TAKS assessment was offered both on computer and on paper because the size of the mode effect may not be the same from one administration to the next. Mode effects may be due to factors such as test questions, test scoring, testing conditions, and examinee groups (Kolen, 1999), and these factors may differ from one test administration to another. For example, one test form may contain more scrolling items than another test form. Based on previous research, we would expect the form with more scrolling items to show a larger performance difference between computer and paper.

Although an overall mode effect may be small, even a small effect can be meaningful and have significant consequences. At the exit level in Texas, passing the TAKS tests is required for high school graduation. A mode difference of even one point on an exit level test can mean a substantial number of students not passing, and hence not receiving their diplomas, because they took the test in a different mode. Thus, while improvements are continually being made to the computer-based testing interface in order to mitigate mode effects, the high-stakes impact of these tests emphasizes the need for regular comparability studies in Texas. If mode differences are found, then scores should be adjusted to provide fairness to students and other test stakeholders. Thus, based on the findings in this literature review, it is recommended that comparability studies continue to be conducted to help ensure the defensibility of the testing programs in Texas and in other statewide assessment programs.

References

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

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

Applegate, B. (1993). Construction of geometric analogy problems by young children in a computer-based test. *Journal of Educational Computing Research, 9*(1), 61-77.

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

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6*(9). Available from http://www.jtla.org.

Bergstrom, B.(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.

Breland, H., Lee, Y. W., & Muraki, E.(2005). Comparability of TOEFL CBT essay prompts: response-mode analyses. *Educational and Psychological Measurement, 65* (4), 577-595.

Bridgeman, B., & Cooper, P. (1998). *Comparability of scores on word-processed and handwritten essays on the Graduate Management Admissions Test.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (RR-01-23). Princeton, NJ: Educational Testing Service.

Campus Computing Project. (2000). *The 2000 national survey of information technology in U.S. higher education: Struggling with IT staffing.* Available from http://www.campuscomputing.net/summaries/2000/index.html.

Cerillo, T. L., & Davis, J. A. (2004). *Comparison of paper-based and computer based administrations of high-stakes, high-school graduation tests*. Paper presented at the Annual Meeting of the American Education Research Association, San Diego, CA.

Chin, C. H. L., Donn, J. S., & Conry, R. F. (1991). Effects of computer-based tests on the achievement, anxiety, and attitudes of grade 10 science students. *Educational and Psychological Measurement, 51*(3), 735-745.

Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper and computer-based assessment in a K-12 setting*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Davis, L. L, Strain-Seymour, E., Lin, C., & Kong, X. (2008). *Evaluating the comparability between online and paper assessments of essay writing in the Texas Assessment of Knowledge and Skills.* Presentation given at the Annual Conference of the Association of Test Publishers, Dallas, TX.

Hetter, R. D., Segall, D. O., & Bloxom, B. M. (1994). A comparison of item calibration media in computerized adaptive testing. *Applied Psychological Measurement, 18*(3), 197-204.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning, and Assessment, 3*(4). Available from http://www.jtla.org.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 5*(2). Available from http://www.jtla.org.

Ito, K., & Sykes, R. C. (2004). *Comparability of scores from norm-reference paper-and-pencil and web-based linear tests for grades 4-12*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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, Montreal, Canada.

Johnson, M., & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and Assessment, 4*(5). Available from http://www.jtla.org.

Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas assessment of knowledge and skills. *Applied Measurement in Education, 21*(3), 207-226.

Kingston, N. M. (2002). *Comparability of scores from computer- and paper-based administrations for students in grades K-8*. 32nd annual Large-Scale Assessment Conference of the Council of Chief State School Officers, Palm Desert, CA.

Kim, D.-H, & Huynh, H. (2006). *A comparison of student performance between paper-and-pencil and computer-based testing in four content areas*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Kim, D.-H., & Huynh, H. (2007). Comparability of computer and paper-and-pencil versions of Algebra and Biology assessments. *Journal of Technology, Learning, and Assessment, 6*(4). Available from http://www.jtla.org.

Kolen, M. J. (1999). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment, 6*(2), 73-96.

Michigan Department of Education (2006). *Evaluation of the first year online Michigan Educational Assessment Program administration.* Retrieved from http://www.michigan.gov/oeaa.

Nichols, P., & Kirkpatrick, R. (2005). *Comparability of the computer-administered tests with existing paper-and-pencil tests in reading and mathematics tests*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

O'Malley, K. J., Kirkpatrick, R., Sherwood, W., Burdick, H. J., Hsieh, M. C., & Sanford, E. E. (2005). *Comparability of a paper based and computer based reading test in early elementary grades*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Oregon Department of Education (2007). *Comparability of student scores obtained from paper and computer administrations.* Retrieved from http://www.ode.state.or.us.

Paek, P. (2005). *Recent trends in comparability studies.* (Pearson Educational Measurement Research Report 05-05). Available from http://www.pearsonedmeasurement.com/research/research.htm

Pearson. (2002). *Final report on the comparability of computer-delivered and paper tests for Algebra I, Earth Science and English*. Austin, TX: Author.

Pearson. (2003). *Virginia standards of learning web-based assessments comparability study report – Spring 2002 administration: Online & paper tests*. Austin, TX: Author.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer.

Piccaino, A. G., & Seaman, J. (2007). *K-12 online learning: A survey of U.S. school district administrators*. Needham, MA: The Sloan Consortium. Available at http://www.sloan-c.org/publications/survey/pdf/K-12_Online_Learning.pdf

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment,3*(6). Available from http://www.jtla.org.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2*(6). Available from http://www.jtla.org

Pommerich, M. (2007). The effect of using item parameters calibrated from paper administrations in computer adaptive test administrations. *Journal of Technology, Learning, and Assessment, 5*(7). Available from http://www.jtla.org.

Pomplun, M., Frey, S., & Becker, D.F. (2002). The score equivalence of paper and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*(2), 337-354.

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1992). *Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays.* (RR-92-45). Princeton, NJ: Educational Testing Service.

Puhan, G., Boughton, K., & Kim S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing. *Journal of Technology,Learning, and Assessment, 6*(3). Available from http://www.jtla.org.

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

Rosevear, P. D. (n.d.) *The ivy league explores online learning.* Available at http://encarta.msn.com/encnet/departments/elearning/?article=ivyleagueonline.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20). Available at http://epaa.asu.edu/epaa/v7n20.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5*(3). Available at http://epaa.asu.edu/epaa/v5n3.html.

Sandene, B., Horkay, N., Bennett, R., 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.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement*, *26*, 261-271.

Sweedler-Brown, C. (1991). Computers and assessment: The effect of typing versus handwriting on the holistic scoring of essays. *Research & Teaching in Developmental Education*, *8*(1), 5–14.

U.S. Department of Commerce. (2002). *A nation online: How Americans are expanding their use of the Internet.* Washington, DC: Author. Available at http://www.ntia.doc.gov/ntiahome/dn/nationonline_020502.htm.

Wang, S. (2004). *Online or paper: Does delivery affect results? Administration mode comparability study for Stanford Diagnostic Reading and Mathematics tests.* San Antonio, Texas: Harcourt.

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

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

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38,* 19-49.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006, April). *Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA. Available at http://www.pearsonedmeasurement.com/downloads/research/RR_06_01.pdf.

Way, W. D., Davis, L. L. & Strain-Seymour, E. (2008, April). *The validity case for assessing direct writing by computer.* Available from http://www.pearsonedmeasurement.com/news/whitepapers.htm.

Way, W. D., Lin, C., & Kong, J. (2008, March). *Maintaining score equivalence as tests transition online: Issues, approaches, and trends.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Way, W. D., Um, K., Lin, C., & McClarty, K. L. (2007, April). *An evaluation of a matched samples method for assessing the comparability of online and paper test performance.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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