Recent Trends in Comparability Studies

Pamela Paek

Pearson Educational Measurement

August 2005

rr0505

Using testing and assessment to promote learning

Pearson Educational Measurement (PEM) is the largest comprehensive provider of educational assessment products, services and solutions. As a pioneer in educational measurement, PEM has been a trusted partner in district, state and national assessments for more than 50 years. PEM helps educators and parents use testing and assessment to promote learning and academic achievement.

PEM Research Reports provide dissemination of PEM research and assessment-related articles prior to publication. PEM reports in .pdf format may be obtained at:

http://www.pearsonedmeasurement.com/research/research.htm

Abstract

The purpose of this paper is to review the research addressing the comparability of computer-delivered tests and pencil-and-paper tests, and particularly the research since 1993. The first part of this paper summarizes the state of online testing technology and the different methods used in the comparability studies. The second part discusses the results from the studies, specifically in K-12 testing. The last part discusses the potential of online assessments.

Research in K-12 education shows that students are using computers in school for classroom-based activities as part of their everyday learning (U.S. Department of Commerce, 2002). In addition, the disparity in computer access among K-12 students has been shown to be negligible over the past five years (Kleiner & Lewis, 2003; Parsad, Jones, & Greene, 2005). Studies indicate that K-12 students are familiar with computers and feel comfortable using them (DeBell & Chapman, 2003; Higgens, Russell & Hoffmann, 2005; Poggio, Glasnapp, Yang, & Dunham 2005; O'Malley et al., 2005; Ito & Sykes, 2004).

With respect to specific comparability research, evidence has accumulated to the point that it appears that the computer may be used to administer tests in many traditional multiple-choice test settings without any significant effect on student performance. One exception to this finding is with respect to tests that include extensive reading passages; for these, studies have tended to show lower performance on computer-based tests than on paper tests (Mazzeo & Harvey, 1988; Murphy, Long, Holleran, & Esterly, 2000; Choi & Tinkler, 2002; O'Malley, Kirkpatrick, Sherwood, Burdick, Hsieh, & Sanford, 2005). These differences may be due to issues related to scrolling and the strategies that students use to organize information (e.g., underlining key phrases). As students continue to become more familiar with reading on the

computer and as computer interfaces begin to include tools to enhance student's reading comprehension, these differences may disappear.

Introduction

The purpose of this paper is to review the research addressing the comparability of computer-delivered and paper-and-pencil tests, and particularly, the research since 1993. The first part of this paper summarizes the state of online testing technology and the different methods used in comparability studies. The second part discusses the results from the studies, specifically in K-12 testing. The last part discusses the potential of online assessments.

There are several reasons for moving to computerized assessments: they can minimize the lag time in score reporting; they can allow for analysis of student performance that cannot be studied from paper tests alone (e.g. individualized assessment tailored to student needs); and they can reduce the consumption of paper and the physical and cost burden of mailing materials (Bennett, 2002b; 2003; Olson, 2003). Given the timeframe for reporting scores for No Child Left Behind (NCLB), computer-based tests seem like a logical choice for the K-12 setting, allowing for instant scoring and reporting of results. On the one hand, rapid scoring and reporting would allow schools to administer tests a few weeks later in the school year, and thus to cover more of the academic curricula that otherwise is cut short by large-scale testing in March and April. On the other, the elimination of scoring lag time will give teachers valuable information to help guide their instruction for the remainder of the school year.

However, the transition from paper-and-pencil to computerized tests cannot be taken for granted. Comparability between the two testing modes must be established through carefully designed and conducted empirical research. This is why comparability studies exist. Comparability studies explore the possibility of differential effects due to the use of computer-based tests instead of paper-and-pencils tests These studies help ensure that test score interpretations remain valid and that students are not disadvantaged in any way by taking a

computerized test instead of the typical paper test. For example, the American Psychological Association's *Guidelines for Computer-Based Tests and Interpretations* (1986) states: "When interpreting scores from the computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cut scores obtained from conventional tests." (p. 18). The joint *Standards for Educational and Psychological Testing* (AERA, 1999) recommends empirical validation of the computerized versions of tests: "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." (p. 57).

Comparability studies analyze the effects on student performance of the mode of test delivery, i.e., either computer- or paper-based testing; these are referred to as "mode effects." There are several avenues to explore in terms of possible mode effects. First, it is important to investigate whether the computer introduces something unintended into the testing situation (Harmes & Parshall, 2000; Parshall, Spray, Kalohn, & Davey, 2002). For instance, researchers should determine how closely computer expertise is related to student performance; since the intent of the computerized test is to assess a construct (e.g. student achievement) other than computer expertise (Harmes & Parshall, 2000; Parshall, et al., 2002), it is hoped that this relation will be negligible. Second, computer and paper test-taking strategies may differ across modes (Murphy, Long, Holleran, & Esterly (2000), which may impact performance in the two testing modes. Third, states that test in the K-12 arena are concerned about how mode effects may differ across academic subjects, item types, and different examinee groups (Hamilton, Klein, and Lorié, 2000). All of the above comprise the primary issues of comparability research.

Some of the early research suggested that electronic page turners, or mere computerized versions of paper tests, are more difficult than the actual paper tests (Bunderson, Inouye, &

Olsen, 1989; Mazzeo & Harvey, 1988; Mead & Drasgow, 1993). One factor contributing to this difficulty was the novelty of the design: students had to learn how to navigate an unfamiliar interface while taking a high-stakes test. Another major issue for students in the early comparability studies was the inability to move back and forth between questions, which students sometimes do on paper tests. Today, however, computer tests routinely permit students to navigate through all items within a section as they would be able to do on paper tests (Pommerich & Burden, 2000).

As Parshall, et al.(2002) indicate, the phrase "computer-based test" has become a catch-all phrase that includes all kinds of computerized assessments (e.g. electronic page turners, computerized tests, computer-adaptive tests). They define computer-fixed tests as fixed both in terms of test form and length. Computer-fixed tests that contain the same items as their corresponding paper tests are also known as electronic page-turners. An advantage of a computer-fixed test is that it mimics the paper test as much as possible, thus limiting the number of confounding factors when comparing test results in the two modes. For instance, current computer-fixed tests allow students to skip items, as well as change answers on previous items, like they may do on a paper test, which keeps possible test-taking strategies consistent across modes. One disadvantage of the computer-fixed test is that students must still encounter items that are too easy or too difficult, since the form is fixed in length and can potentially cover a large range of item difficulties. Computer-fixed tests also take little advantage of the special capabilities of computers In addition, test security can be difficult to maintain with computer-fixed tests since the number of available test forms is usually limited.

One way around the computer-fixed test is a computer-adaptive test (CAT). In such a test, students are administered items based on the ability level they demonstrate on early

questions, rather than on a fixed set of items administered to all students, which are likely to contain items that are either too difficult or too easy for students. The assumption is that CATs provide a more accurate and efficient measure of student performance. The main advantage of computer-adaptive tests is the potential for decreased testing time, as CATs can hone in on student ability in fewer items than a computer-fixed test or paper test. A disadvantage of most computer-adaptive tests is that students cannot navigate freely through the test as they would a computer-fixed or a paper test: generally, students can only progress forward on computer-adaptive tests. This difference can affect motivation. It may also result in the assessment of a different set of test-taking skills from those assessed in a paper test or computer-fixed test. The degree to which a CAT can reduce testing time or improve measurement accuracy is dependent on the breadth of the construct domain and the depth of the item pool.

Wang and Kolen (2001) discuss several reasons why establishing the comparability of computer-adaptive tests and paper tests is both challenging and necessary. On the one hand, constraining a CAT to be comparable to a paper test restricts the advantages that computer-adaptive tests offer. Yet, if both computer-adaptive and paper versions of the same test are to be used, professional standards suggest that reported scores still be comparable. Kolen (1999-2000) summarized four aspects of possible mode differences: (1) differences in test questions, (2) differences in test scoring, (3) differences in testing conditions, and (4) differences in examinee groups. For instance, the scoring methods for computer-adaptive tests differ from those for paper tests because different sets of items are administered to different examinees in CAT. Wang and Kolen (2001) suggested that using ability estimates based on item response theory (IRT) as part of computer-adaptive tests may exacerbate differences between computer-adaptive and paper versions of a test. Stocking (1997) explored the utility of using the number-correct raw score

instead of IRT ability estimates when scoring computer adaptive tests using the three-parameter logistic model. She found that the accuracy of the two scoring methods in this context were surprisingly similar. However, she did not compare the two computer-adaptive test scoring methods in terms of comparability with paper test scores. Wang and Kolen (2001) also found that changes in components of computer-adaptive tests, such as exposure rate control, content balancing, test length, and item-pool size could result in different levels of comparability between computer-adaptive and paper test scores.

Methodologies in Comparability Studies

A majority of the comparability research has focused on the differences in means and standard deviations of test scores (e.g. Makiney, Rosen, & Davis, 2003; Mead & Drasgow, 1993; Merten, 1996; Pinsoneault, 1996) with little focus on underlying measurement issues (Donovan, Drasgow, & Probst, 2000; King and Miles, 1995). Raju, Laffitte, and Byrne (2002) state that "without measurement equivalence, it is difficult to interpret observed mean score differences meaningfully."

As such, there are several ways to compare performance across testing modes. For instance, there is the root mean-squared difference (RMSD) statistic, which provides an estimate of the average difference in item difficulties due to the assessment mode:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}\left({}_{\mathrm{CBT}}\hat{b}_i - {}_{PPT}\hat{b}_i\right)^2}{n}}.$$

There are also classical item and form analyses: p-values, response-option (distracter) analyses, item-to-total-score correlations, item difficulty values, item-model fit measures, and standard errors of measurement. Traditional forms analyses include test form reliability, summary descriptive statistics, frequency distributions and percentile rank information. A

comparison of person-fit statistics can also be made across modes, and can be disaggregated by various breakout groups including ethnicity and gender.

Makiney, et.al. (2003) discuss two common techniques for examining measurement equivalence: (a) structural equation modeling methods involving confirmatory factor analysis and (b) methods based on item response theory (Raju, Laffitte, & Byrne, 2002). Confirmatory factor analysis may be preceded by use of an exploratory principal factor analysis. Tests for invariance may also accompany confirmatory factor analyses, such as analyzing for measurement equivalence through equivalence/invariance and equality of the number of factors and factor pattern matrices. Raju, et al. (2002) recommend using confirmatory factor analysis when analyzing multiple latent constructs and multiple populations simultaneously.

In terms of item response theory (IRT), if the model assumptions are satisfied, measurement equivalence is considered satisfied whenever the item parameters are equal across the two subpopulations (Raju, et al., 2002). There are two types of IRT analysis to test for the equivalency of item parameters. The first, differential item functioning analysis (DIF) compares performance on individual items between the two modes while still controlling for overall mode difficulty. The second, differential test functioning, differs from the traditional IRT DIF technique by examining the test as a whole, rather than individual items (Makiney, et al., 2003). Raju, et.al. (2002) recommend using the IRT method for multiple-choice items, since a logistic regression model is considered appropriate for expressing the relationship between a continuous underlying construct and a measured variable.

Comparability Studies Prior to 1993

Bunderson, Inouye, and Olsen (1989), Mazzeo and Harvey (1988), and Mead and Drasgow (1993) collectively analyzed more than 89 comparability studies prior to 1993. These

studies utilized a variety of tests (e.g. aptitude tests, personality tests, achievement tests), mainly with adults as subjects. Bunderson, et al. (1989) analyzed 23 comparability studies: three studies showed students performing better on the computer tests; eleven showed no significant differences in mode; and studies showed students performing better on the paper tests. Of the eleven studies that showed the two modes to be comparable, the paper test scores were found to be slightly higher than the computer test scores.

Mazzeo and Harvey (1988) conducted a literature review of 38 studies (44 tests) to better understand what effects the computer may have on the psychometric aspects of testing: eleven studies showed students performing better on the computer test, eighteen showed no differences in mode, and fifteen showed students performing better on the paper tests. They analyzed the results by type of test administered (e.g. psychological scales, general aptitude tests, achievement tests, speed tests, timed power tests). One finding was that students omitted a different number of items on paper than on the computer. Also, graphical display tended to have an effect on comparability. Tests with reading passages appeared more difficult for students when administered on the computer. They concluded that although some studies showed no mode differences, paper tests appeared easier than computer versions of the same test. They hypothesized that one reason for the lower scores for computer-based tests may be accidental keying or clicking of the wrong answer.

Mead and Drasgow (1993) used meta-analytic techniques to analyze 28 studies (159 tests) comparing performance of young adults and adults on computerized and paper assessments. Of these 28 studies, 11 used the Armed Services Vocational Aptitude Battery (ASVAB), comprising 59 tests; two used the Differential Aptitude Test (DAT), comprising 16 tests; one used the Graduate Research Examination (GRE), comprising three tests; two used the

Multidimensional Aptitude Battery (MAB), comprising 10 tests; and one used the Western Personnel Test (WPT), comprising two tests. The remaining studies did not use a standardized battery. Almost half of the studies used military recruits for their samples.

They found that there was no evidence of mode differences on power tests, which are idealized as content-focused, untimed tests. They also did not find significant differences between computer-adaptive tests and computer-fixed tests. However, they did find a significant mode effect for timed (speeded) tests. One limitation of the Mead and Drasgow meta-analysis is that the researchers did not separately evaluate mode differences across content areas.

Technology Advancements Since 1993

After Mead and Drasgow published their article, the amount of computer use began to change. In 1993, only 27% of students used computers (DeBell & Chapman, 2003); the lack of familiarity with computers may have resulted in lower computer test performance in these earlier studies. In the latest surveys of computer use, 90% of students ages 5-17 use the computer (DeBell & Chapman, 2003); in September 2001, 75% of students ages 5-9 and 85% ages 10-17 use computers in the classroom (U.S. Department of Commerce, 2002). In addition, use of the Internet in public schools has greatly increased since 1994. In 1994, only 3% of classrooms had access to the Internet, usually through dial-up connection; in 2003, 93% of classrooms had Internet access through broadband connection. In 2003, all public schools reported having Internet access, regardless of school demographics (e.g. school size, percent minority enrollment, percent of students eligible for free or reduced-price lunch), compared to only 35% of public schools in 1994 (Parsad, et al., 2005). Internet access at the school level has remained consistent since 1999 (Kleiner & Lewis, 2003; Parsad, et al., 2005). Given the prevalence of computers and the Internet in schools and at homes, it appears that now, more than ever, technology can be used

for assessing student performance and that differences in student performance on computer and paper tests can be expected to diminish.

The Increase in Comparability Studies

Since Mead and Drasgow, the number of comparability studies has increased, mainly due to increased school access to computers and the Internet. Some of these studies have followed Mead and Drasgow's recommended next steps and have looked specifically at certain factors that can be attributed to differential performance by mode.

Mead and Drasgow (1993) isolated two factors that they hypothesized could contribute to potential differences between paper test and computer tests: the possible effect of scrolling long reading passages on the computer and the graphical display of questions or tasks on the computer. These factors have been studied since the publication of Mead and Drasgow's article, and in fact, changes to technology have allowed for some advancement in the study of these specific factors on mode effects. For instance, several studies have indicated that when all the information for an item is presented entirely on the screen, mode effects are insignificant (Bridgeman, Lennon, & Jackenthal, 2001; Choi & Tinkler, 2002; Hetter, Segall, & Bloxom, 1994). Other studies have indicated significant mode effects when students must scroll or navigate through information on the computer screen to answer test questions (Bridgeman, et al., 2003; Choi & Tinkler, 2002). However, these findings have not been universally accepted. For example Pomplun, Frey and Becker (2002) asserted that difficulties in reading on computer screens were related to primitive technology.

Other results from comparability studies show that the time to complete computer tests is significantly shorter than the time to complete paper tests (Alexander, Bartlett, Truell, & Ouwenga, 2001; Bunderson, et.al., 1989; Pomplun, Frey, & Becker, 2002; van de Vijver &

Harsveld, 1994; Wise & Plake 1989). This may be due to the simplicity of clicking an answer choice with a mouse or typing an answer from the keyboard, rather than bubbling an answer on a separate answer document. For multiple-choice tests, the research suggests that differences in computer experience have virtually no effect on test scores (Bennett, 2002b; Bridgeman, et al., 1999; Taylor, et al., 1998). In addition, several studies have found that demographic differences are not significantly related to differences in performance on computer tests and paper tests (Alexander, et al., 2001; Nichols & Kirkpatrick, 2005).

All of the studies above have honed in on different issues in comparability studies. K-12 assessment is yet another arena in studying the comparability of computer and paper testing. The remainder of this paper focuses on the results of K-12 comparability studies.

K-12 Comparability Studies

Comparability studies in K-12 prior to 1993 are rare, with small samples of students and mixed findings. For instance, Gerrell and Mason (1983) found that fifth-graders comprehended computer-displayed passages better than printed passages while Zuk (1986) found that third- and fifth-graders had slower comprehension when reading text on the computer. Feldmann and Fish (1988) studied elementary, junior high, and high school students. These researchers discovered that lack of computer access was not related to the ease of reading text on the computer in elementary and junior high students. They reported that the ability to comprehend text on the computer and paper were the same. Applegate (1993) found that kindergarteners were not capable of constructing geometric analogy problems on the computer; they performed significantly worse on computer than on paper and made errors that were not made by the second-grade students that were also assessed.

Most K-12 comparability studies have been conducted within the past few years, and results tend to show computer tests to be equivalent in difficulty or slightly easier than paper tests (Bridgeman, Bejar, & Friedman, 1999; Choi & Tinkler, 2002; Hetter, et al., 1994; Pearson Educational Measurement, 2002; 2003; Poggio, Glasnapp, Yang, and Poggio, 2005; Pommerich, 2004; Pomplun, 2002; Russell, 1999; Russell & Haney, 1997; 2000; Russell & Plati, 2001; Taylor, Jamieson, Eignor, & Kirsch, 1998; Zandvliet & Farragher, 1997; Wang, 2004). Where mode differences were detected, they were not always statistically significant. There may be two reasons for comparability in computer and paper test performance. One, computer access and familiarity are less of a concern today than in previous years. Two, many computer tests now have tools that give students more freedom, allowing them to skip items and come back to them, and to change their answers. These new navigational tools simulate the test-taking strategies students may use when taking a paper test, resulting in more equivalent scores (Mason, Patry, & Bernstein, 2001; Wise & Plake, 1989). They were not available in the earlier years of computer tests. At that time, computer test-taking strategies most likely differed from those used for paper tests. For instance, Chin, et al. (1991) found that tenth-grade science students performed better on the computer test. They hypothesized that guessing was not as common on the computer test as on the paper test. Chin and her colleagues thought that students tried harder on the computer test, because students tended to avoid the response "I don't know" and thus may have changed their tendency to guess. As long as the items were sufficiently simple, computer tests such as electronic page turners without navigational functionality proved comparable to paper tests, even though a different test-taking strategy was used on the computer test (e.g., with each item administered individually, students were not able to review other responses and possibly get cues from other items).

Recent studies show comparable results for computer and paper tests across subjects in K-12, regardless of test-taking strategies. For example, Pommerich (2004) analyzed students in eleventh and twelfth grade on a computer-fixed test in English, reading, and science reasoning in 1998 and 2000. In 1998, the difference between the paper and computer-fixed test was about one raw score point: one point higher for the computer version of the English test, and one point higher for the paper version of the reading test. In 2000, there was less than one raw score point difference for reading and science, but the computer scores were more than one raw score point higher in English. In her study, students performed better on questions near the end of the computerized tests than on the same items on the paper version. Pommerich postulated that the computer makes it easier to respond and move quickly through items, thus helping students needing more time at the end of the test.

Pomplun, Frey, and Becker (2002) had high school, community college, and four-year university students take two forms of the Nelson-Denny Reading Test, which is a speeded reading comprehension test. Regardless of whether the student took the paper test or the computer test first, the scores were higher on the computer tests. One sample had a significantly higher mean score on the computer test, about seven points higher. These high scores seem related to the difference of using a mouse rather than a separate answer sheet to record responses. Students finished the computer test more quickly than the paper test. In addition, the completion rate was higher for the computer test than the paper test, which may be due to the fact that students tend to omit items less often on computer tests than paper tests, even if the opportunity is available.

Russell (1999) found that open-ended questions can result in differences by mode. Grade 8 students were tested in science, language arts, and math open-ended items (6 items per

subject). For math, students performed, on average, one point higher on the paper test, but this was not a statistically significant difference. There was a statistically significant difference for science, with students scoring on average more than two points better on the computer test. Language arts performance was the same for the paper test and the computer test.

Several studies have found that students tend to write more on computer tests than on paper tests, but that the writing is not necessarily of better quality (Dauite, 1986; Etchison, 1989; Nichols, 1996; Russell, 1999; Russell & Haney, 1997). In fact, the quality of writing appears the same whether written on paper or on computer for middle school and high school students (Dauite, 1986; Etchison, 1989; Grejda, 1989; Hawisher & Fortune, 1989; Nichols, 1996; Russell, 1999; Russell & Haney, 1997).

Russell (2002) and Russell and Haney (1997; 2000) found that middle school students (grades 6, 7, 8 in math, science and language arts) perform similarly on paper tests and computer tests for multiple-choice questions and significantly better on computer tests for science and language arts with short answers. For students familiar with writing on computer, writing scores were significantly higher than on paper. The effect size was found to be 0.94 on the extended writing task and 0.99 and 1.25 for the NAEP language arts and science short-answer items. For the extended writing task, students' writing appeared to be underestimated when tested by paper test as compared to computer test, with effect sizes of 0.4 to 1.0. These effect sizes have both statistical and practical significance (Cohen, 1977; Wolf, 1986), as an effect size of 0.94 indicates that the average student in the computer mode performed better than 83% of students in the paper mode.

A comparability study using the National Assessment of Educational Progress (NAEP) for students in eighth grade was also conducted. R. Bennett (personnel communication,

December 16, 2004) reported no significant differences in the mean scores of grade 8 students performing online and paper writing tasks as part of the NAEP Writing Online study.

Wang (2004) studied performance on the Stanford Diagnostic Reading and Mathematics Tests, fourth edition (SDRT 4 and SDMT 4), for students in grades 2-12. Overall, there were no significant differences in total test score means based on administration mode, mode order, or mode-by-mode order interactions. However for level 6 (grades 9-12), the differences in the means based on administration mode and mode order were statistically significant for reading, and the differences in the means based on mode-by-mode order interactions were statistically significant for math. Wang did not indicate whether these differences were in favor of the paper or the online version.

However, some current K-12 studies are also finding students to be performing worse on computer tests than paper tests (Cerrillo & Davis, 2004; O'Malley, Kirkpatrick, Sherwood, Burdick, Hsieh, & Sanford, 2005). In one case, there were also mitigating factors in the design of the study, which may have influenced comparability. Cerillo and Davis (2004) discovered that students performed 4-7% better on the paper version of a high school graduation test compared with a computerized version. Alone, this information may imply that students are still not practiced enough in their use of computers to be assessed by computer tests. Unfortunately, in this study, only the paper tests counted toward graduation, so the incentive to perform on the computer tests was not similar to that on the paper tests. A common way to address this limitation in comparability studies is to allow students to choose the better of the two performances. This helps to motivate students to try equally hard on both versions of the test (Cerillo & Davis, 2004; Fitzpatrick & Triscari, 2005; Kiplinger & Linn, 1996; O' Neil, Sugrue, & Baker, 1996; PEM, 2001; 2002; 2003).

The studies reviewed thus far have used the K-12 population; however, they have generally not addressed comparability in the context of high-stakes, large-scale assessment. Klein and Hamilton (1999) state that "Advances in psychometrics and information technology are likely to accelerate the adoption of this [computer test] approach, particularly when the tests are administered (or downloaded into the school) via the Internet." As such, an increasing number of states (Arkansas, Georgia, Idaho, Kansas, Kentucky, Maryland, North Carolina, Oregon, South Dakota, Texas, Utah, and Virginia) are pursuing comparability studies or have already implemented computer tests as part of their K-12 assessment programs (Bennett, 2002b; 2003; Olson, 2003).

Choi and Tinkler (2002) looked at student performance on grade 3 and 10 math and reading of the Oregon state assessment. They found that the computer tests in reading were more difficult than the paper tests, especially for third-graders. They noted that as students age, they acquire more computer experience; thus, the novelty of being tested via computer is reduced. Choi and Tinkler found that computer familiarity was related to computer test performance, as students who rarely used a computer tended to perform lower in both mathematics and reading than those students who had more computer experience. Students were asked which format they preferred, and for both grades, students tended to prefer the computer test platform. Choi and Tinkler found no significant differences in the mean item difficulties for tenth graders on the math test.

In the Virginia state assessment, students took a common set of items on paper, then a second set of items for which students were randomly assigned to one of two groups (computer fixed or paper test) in English, Algebra I, or Earth Science (Pearson Educational Measurement, 2002). Scores appeared comparable across classical and IRT statistics. Students scored similarly

in the two modes: on average, one raw score point higher on the paper test in both English and Earth Science. In Algebra I, Algebra II and Biology, on average, there was no difference in raw scores or scale scores for students classified in proficient or advanced levels of performance (Pearson Educational Measurement, 2002; 2003). Similarly, in the New York state assessment, eighth-graders were administered an intermediate-level technology examination. Results showed that the computer test was easier by five raw score points. This difference was not statistically significant, as seen with a test statistic of 0.4579 with 304 degrees of freedom using Student's independent t-test and assuming unequal variances (Pearson Educational Measurement, 2001).

Pearson Educational Measurement (2001) compared the performance of eighth-graders taking a paper test with those taking the test online using an Intermediate Level Technology examination for the state of New York. This study included both classical and IRT analyses, examining item-level statistics both overall and by testing mode to determine the quality of the item performance, while simultaneously examining differences between paper-based and computer-based performance. By treating the computer mode as the focus group and the paper testing condition as the reference group, traditional Mantel-Haenszel as well as Rasch IRT DIF analyses were performed. Some items exhibited DIF across modes, but students tended to perform better on the computer version of the test overall.

Poggio, et al. (2005) looked at the performance of seventh grade students on the Kansas state assessment in mathematics. The average difference between computer and paper test scores was less than 1 percentage point. Performance was similar across demographics (gender, academic placement, and SES), with students performing slightly better on the computer test than paper test (by at most 1.5% points).

Nichols and Kirkpatrick (2005) analyzed high school students in grade 11 on Florida's state assessment in mathematics and reading. Tests of mode effects showed no statistically significant differences. The average difference was less than one raw score point in each subject. Performance varied across ethnicity in both subjects, but the authors propose this difference to be due to the samples not being comparable in ethnicity across the two mode groups.

Conclusions

Research in K-12 shows that students are using computers in school for classroom-based activities; they are using current technology as part of their everyday learning (U.S. Department of Commerce, 2002). In addition, the disparity in computer access among K-12 students has been shown to be negligible over the past five years (Kleiner & Lewis, 2003; Parsad et al., 2005). Studies show that K-12 students are familiar with computers and are comfortable using them (DeBell & Chapman, 2003). It appears that we are now able to assess students with current technology and can anticipate comparable results between computer-fixed and paper tests.

The K-12 comparability studies to date show that, in general, computer and paper versions of traditional multiple-choice tests are comparable across grades and academic subjects. Table 1 below summarizes the results of a number of studies by subject and types of tests used. The general findings of comparability across modes may be due to the fact that modern computer test systems allow students to navigate the computer tests as they would a paper test, thus allowing for similar test-taking strategies across modes. Also, the prevalence of computers and the Internet in today's K-12 classrooms allows students to feel comfortable using computers as part of their learning. Collectively, evidence has accumulated to the point where it appears that in many traditional multiple-choice test settings, the computer may be used as a medium to administer tests without any significant effect on student performance.

There may still be one area where there differences remain: items relating to long reading passages (Murphy, Long, Holleran, & Esterly, 2000; O'Malley, et al., 2005). Although issues with text display and scrolling have been studied and adjusted, it appears that tests with extended reading passages remain more difficult on computer than on paper. This appears to be due to the fact that testing on computers may inhibit students' reading comprehension strategies, such as underlining phrases as they would on paper, and that they must scroll down the computer screen to read the entire passage. Issues with scrolling may have to do with visual learning, as on a paper test, students can generalize about where certain information may be located in a passage, as the passage is fixed on a page. With scrolling, students loses that sense of placement. However, there is promise in this area, as new tools are being developed to mimic the way students use reading passages on paper, such as a computerized highlighter. It may be that once students become familiar with such tools that differences found with tests involving long reading passages across modes will be minimized.

Future Steps in Computerized Assessments

This paper has discussed the comparability of computer- and paper-based tests, as that appears to be the first step in transitioning into using technology for assessments. There appear to be several next steps in the transition to computerized testing. First, states should continue to conduct comparability studies of their high-stakes, large-scale assessments across grades and subjects. As we hone in on mode differences by grade or subject, we may be able to recommend improvements in technology to minimize or remove the disparity across modes, as is being done for reading comprehension items. Second, research must continue across different item types, specifically open-ended item types and innovative computerized item types, to better understand the differences that may exist across modes. Third, technology that allows for more complex

types of assessment should be considered, in anticipation of the day when all testing is done by computer and comparability may no longer be an issue.

As Bennett (2002a) indicates, "K-12 agencies have educational responsibilities that may force them to go beyond the initial achievement of computerization to create assessments that support learning and instruction in way that paper test cannot." (p. 14-15).

For instance, the computer is capable of delivering innovative items representing new ways of assessing student learning that can not be replicated in paper testing (Bennett & Bejar, 1998; Perlman, Berger, & Tyler, 1993). For instance, performance-based tasks may be easily simulated on the computer, which allows assessors to better understand what students know within a specific task. In such ways, the full potential of computers to enhance assessment practices can begin to be fulfilled.

Table 1
Results of Comparability Studies in K-12

	Computer Test More Difficult	Paper Test More Difficult	Comparable
Math	• Applegate (1993): Kindergarten • Choi & Tinkler (2002): grade 3 • Cerillo and Davis (2004): Algebra	• Choi & Tinkler (2002): grade 10	• Pearson Educational Measurement (PEM) (2002): Algebra • PEM (2003): Algebra II • Nichols and Kirkpatrick (2005) • Poggio, Glasnapp, Yang, and Poggio (2005): grade 7 • Russell (1999): grade 8 • Russell (2002): grades 6, 7, 8 (MC) • Russell & Haney (1997; 2000) : grades 6, 7, 8 (MC) • Wang (2004), grade 2-5, 7-12
Language Arts		• Russell (2002): grades 6, 7, 8 (MC) • Russell & Haney (1997; 2000) : grades 6, 7, 8 (MC)	• Pommerich (2004): grades 11-12 • Russell (1999): grade 8 • Russell (2002): grades 6, 7, 8 (MC) • Russell & Haney (1997; 2000) : grades 6, 7, 8 (MC)
Reading	• Choi & Tinkler (2002): grade 3 • Cerillo and Davis (2004): English	• Choi & Tinkler (2002): grade 10 • Pomplun, Frey, and Becker (2002): high school • O'Malley et al. (2005), grades 2-5, 8 (MC)	• Nichols and Kirkpatrick (2005) • PEM (2002): English • Pommerich (2004): grades 11-12 • Russell (1999): grade 8 • Russell (2002): grades 6, 7, 8 (MC) • Russell & Haney (1997; 2000) : grades 6, 7, 8 (MC) • Wang (2004), grade 2-5, 7-12
Science	• Cerillo and Davis (2004): Biology	• Chin , Donn, and Conry (1991): grade 10 • Russell (1999): grade 8 • Russell (2002): grades 6, 7, 8 (MC) • Russell & Haney (1997; 2000) : grades 6, 7, 8 (MC)	• PEM (2002): Earth Science • PEM (2003): Biology • Pommerich (2004): grades 11-12 • Russell (2002): grades 6, 7, 8 (MC) • Russell & Haney (1997; 2000) : grades 6, 7, 8 (MC)

References

Alexander, M.W., Bartlett, J.E., Truell, A.D. & Ouwenga, K. (2001). Testing in a computer technology course: An investigation of equivalency in performance between online and paper methods. *Journal of Career and Technical Education, 18*(1), 69-80.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999*), 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: American Psychological Association.

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

Bennett, R. (December 16, 2004). Personal communication.

Bennett, R.E. (2002a). Inexorable and inevitable: The continuing story of technology and assessment. *The Journal of Technology, Learning and Assessment, 1*(1), 1-24.

Bennett, R.E. (2002b). *Using electronic assessment to measure student performance.* The State Education Standard, Washington, DC: National State Boards of Education. Retrieved February 1, 2005, from http://www.nasbe.org/Standard/10_Summer2002/bennett.pdf.

Bennett, R.E. (2003). *Online assessment and the comparability of score meaning.* Princeton, NJ: Educational Testing Service.

Bennett, R.E., & Bejar, I.I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement Issues and Practice,* 17, 9-17.

Bergstrom, B. (1992, April). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis*. Paper presented at the annual meeting of the American Educational Research Association: San Francisco.

Bridgeman, B., Bejar, I. I., & Friedman, D. (1999). Fairness issues in a computer-based architectural licensure examination. *Computers in Human Behavior,* 15, 419–440.

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

Bunderson, C. V., Inouye, D.K., & Olsen, J.B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 367-407). London: Collier Macmillan.

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.

Clansing, C., & Schmitt, D. (1990). *Paper versus CRT: Are reading rate and comprehension affected?* In Proceedings of selected paper presentation at the convention of the Association for Educational Communications and Technology. (ERIC Document Reproduction Service No. ED323924).

Cohen, J (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.) NY: Academic Press.

Dauite, C. (1986). Physical and cognitive factors in revising: insights from studies with computers. *Research in the Teaching of English*, 20, 141-159.

DeBell, M. & Chapman, C. (2003). *Computer and Internet use by children and adolescents in 2001: Statistical Analysis Report*. Washington, DC: National Center for Education Statistics.

Donovan, M.A. Drasgow, F., Probst, T.M. (2000).Does computerizing paper-and-pencil job attitude scales make a difference? New IRT analyses offer insight. *Journal of Applied Psychology*, *85*(2), 305–313.

Etchison, C. (1989). Word processing: A helpful tool for basic writers. *Computers and Compostion, 6*(2), 33-43.

Feldmann, S. C., & Fish, M. C. (1988). Reading comprehension of elementary, junior high and high school students on print vs. microcomputer-generated text. *Journal of Educational Computing Research,* 4(2), 159-166.

Fitzpatrick, S., & Triscari, R. (2005, April). *Comparability Studies of the Virginia computer-delivered tests*. Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada.

Gerrell, H.R., & Mason, G.E. (1986). Computer-chunked and traditional text. *Reading World, 22*, 241-246.

Grejda, G.F. (1992). Effects of word processing on sixth grader's holistic writing and revision. *Journal of Educational Research, 85*(3), 144-149.

Hamilton, L.S., Klein, S.P., & Lorie, W. (2000). *Using web-based testing for large-scale assessment*. Santa Monica, CA: RAND Corporation.

Harmes, J.C., & Parshall, C.G. (2000, November). *An iterative process for computerized test development: Integrating usability methods*. Paper presented at the annual meeting of the Florida Educational Research Association, Tallahassee.

Hawisher. G.E., & Fortune, R. (1989). Word processing and the basic writer. *Collegiate Microcoputer, 7*(3), 275-287.

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.

King, W.C., & Miles, E.W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper measurements: A test of measurement equivalence. *Journal of Applied Psychology, 80,* 643-651.

Kiplinger, V.L., & Linn, R.L. (1996). Raising the stakes of test administration: The impact of student performance on the National Assessment of Educational Progrss. *Educational Assessment 3*(2), 111-133.

Klein, S.P., & Hamilton, L. (1999). *Large-scale testing: Current practices and new directions*. Santa Monica, CA: RAND.

Kleiner, A., & Lewis, L. (2003). *Internet access in U.S. public schools and classrooms: 1994-2002* (NCES 2004-011). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Kolen, M.J. (1999-2000). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment 6*, 73-96.

Makiney, J.D., Rosen, C., Davis, B.W., Tinios, K. & Young, P. (2003). *Examining the measurement equivalence of paper and computerized job analyses scales.* Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Mason, B.J., Patry, M. & Bernstein, D.J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. *Journal of Educational Computing Research, 24*(1), 29-39.

Mazzeo, J., & Harvey, A.L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature* (ETS RR-88-21). Princeton, NJ: Educational Testing Service.

Mead, A.D. & Drasgow, F. (1993). Equivalence of computerized and paper cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449-458.

Merten, T. (1996). A comparison of computerized and conventional administration of the German versions of the Eysenck Personality Questionnaire and the Carroll Rating Scale for Depression. *Personality and Individual Differences, 20*, 281-291.

Murphy, P.K., Long, J., Holleran, T., & Esterly, E. (2000, August*). Persuasion online or on paper: A new take on an old issue*. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Neuman, G., & Baydoun, R. (1998). Computerization of paper tests: When are they equivalent? *Applied Psychological Measurement, 22*(1), 71-83.

Nichols, L.M. (1996). Paper and pencil versus word processing: a comparative study of creative writing in the elementary school. *Journal of Research on Computing in Education, 29*(2), 159-166.

Nichols, P. & Kirkpatrick, R. (2005, April). *Comparability of the computer-administered tests with existing paper-and-pencil tests in reading and mathematics tests*. Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada.

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

O'Malley, K. J., Kirkpatrick, R., Sherwood, W.,Burdick, H. J., Hsieh, M.C.,Sanford, E.E. (2005, April). *Comparability of a Paper Based and Computer Based Reading Test in Early Elementary Grades*. Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada.

O'Neil, H.,F., Sugrue, B., & Baker, E.L. (1996). Effects of motivational interventions on the NAEP mathematics performance. *Educational Assessment 3*(2), 135-157.

Palacio-Cayetano, J.; Allen, R. D.; Stevens, R. H. (1999). The American Biology Teacher 61(7), 514-522.

Parsad, B., Jones, J. & Greene, B. (2005). *Internet access in U.S. public schools and classrooms: 1994-2003* (NCES 2005-015) U.S. Department of Education. Washington, DC: National Center for Education Statistics.

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

Perlman, M., Berger, K., & Tyler, L. (1993). *An application of multimedia software to standardized testing in music (*ETS RR 93-36). Princeton, NJ: Educational Testing Service.

Pearson Educational Measurement. (2001). *The comparability of paper-based and online responses to the intermediate-level test of Technology Education, State of New York.* Austin, TX: Author.

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

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

Pinsoneault, T.B. (1996). Equivalency of computer-assisted paper-and-pencil administered version of the Minnesota Multiphasic Personality Inventory-2. *Computers in Human Behavior, 12*, 291-300.

Poggio, J., Glasnapp, D.R., Yang, X. & Poggio, A.J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *The Journal of Technology, Learning, and Assessment, 3*(6), 1-30.

Pommerich, M. (2004). Developing computerized versions of paper tests: Mode effects for passage-based tests. *The Journal of Technology, Learning, and Assessment, 2*(6), 1-44.

Pommerich, M., & Burden, T. (2000, April). From simulation to application: Examinees react to computerized testing. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

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.

Raju, N.S., Laffitte, L.J., & Byrne, B.M. (2002). Measurement equivalence: A comparison of confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87* (3), 517-529.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20). Retrieved February 1, 2005, from 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 . *Educational Policy Analysis Archives, 5*(3). Retrieved February 1, 2005, from http://epaa.asu.edu/epaa/v5n3.html

Russell M. & Haney.W. (2000). Bridging the gap between testing and technology in schools. *Education Policy Analysis Archives, 8*(19). Retrieved February 1, 2005, from http://epaa.asu.edu/epaa/v8n19.html

Russell, M. & Plati, T. (2001). Effects of computer versus paper administration of a state-mandated writing assessment. Teachers College. Retrieved February, 2005, from http://www.tcrecord.org

Stocking, M.L. (1997). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21*, 365-389.

Taylor, C., Jamieson, J., Eignor, D. & Kirsch, I. (1998). The relationship between computer familiarity and performance on computer-based TOEFL test tasks.

U.S. Department of Commerce. (2002). *A nation online: How Americans are expanding their use of the Internet*. Washington, DC: Author.

Van de Vijver, F.J.R., & Harsveld, M. (1994). The incomplete equivalence battery of the paper and computerized versions of the general aptitude test battery. *Journal of Applied Psychology, 79,* 852-859.

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, T. & Kolen, M.J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38*(1), 19-49.

Wise, S.L. & Plake, B.S. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice*, 8(3), 5-10.

Wolf, F. (1986) *Meta-analysis: Quantitative methods for research synthesis*. SAGE University series on quantitative applications in the social sciences, series no. 07-059. Newbury Park, CA: SAGE.

Zuk, D. (1986). The effects of microcomputers on children's attention to reading. *Computers in the Schools, 3,* 39-51.