Comparability of Computer-Based and Paper-and-Pencil Testing in K 12 Reading Assessments: A Meta-Analysis of Testing Mode Effects

Shudong Wang, Hong Jiao, Michael J. Young, Thomas Brooks and John Olson

Educational and Psychological Measurement 2008; 68; 5 originally published online Sep 12, 2007;
DOI: 10.1177/0013164407305592

Educational and Psychological Measurement
Volume 68 Number 1
February 2008 5-24

10.1177/0013164407305592
http://epm.sagepub.com
hosted at
http://online.sagepub.com

Comparability of Computer-Based and Paper-and-Pencil Testing in K–12 Reading Assessments

A Meta-Analysis of Testing Mode Effects

Shudong Wang
Hong Jiao
Michael J. Young
Thomas Brooks
John Olson
Harcourt Assessment, Inc.

In recent years, computer-based testing (CBT) has grown in popularity, is increasingly being implemented across the United States, and will likely become the primary mode for delivering tests in the future. Although CBT offers many advantages over traditional paper-and-pencil testing, assessment experts, researchers, practitioners, and users have expressed concern about the comparability of scores between the two test administration modes. To help provide an answer to this issue, a meta-analysis was conducted to synthesize the administration mode effects of CBTs and paper-and-pencil tests on K–12 student reading assessments. Findings indicate that the administration mode had no statistically significant effect on K–12 student reading achievement scores. Four moderator variables—study design, sample size, computer delivery algorithm, and computer practice—made statistically significant contributions to predicting effect size. Three moderator variables—grade level, type of test, and computer delivery method—did not affect the differences in reading scores between test modes.

Keywords: *meta-analysis; computer-based testing; comparability of educational test modes; K–12 reading tests*

Reading plays a prominent role in K–12 education and students' futures. Reading is the most frequently measured achievement construct (Stenner, 1996) compared to the rest of regular curricula such as mathematics, science, social

Authors' Note: The authors are grateful for the insightful suggestions of two anonymous reviewers. We would also like to express our sincere thanks to the editor for his careful editing and valuable comments. Please address correspondence to Shudong Wang, PhD, Psychometrics and Research Services, Harcourt Assessment, Inc., 19500 Bulverde Road, San Antonio, TX 78259; e-mail: shudong_wang@harcourt.com.

science, and other subjects in K–12 education. It is part of most sets of content standards including those developed by every state as well as the National Assessment of Educational Progress. The importance of reading is also emphasized in the No Child Left Behind (NCLB) Act, which requires content standards, academic achievement standards, and aligned assessments at Grades 3–8 and at the high school level. Many states have had to expand their assessment programs recently to meet these requirements. Reading assessments measure the critical learning area to be a competent citizen. Increasingly, these assessments are being conducted by computer-based delivery systems. As information technology advances, computers have become indispensable to facilitating classroom instruction and assessment, and they are positively received by students and teachers.

The implementation of the NCLB Act has increased the stakes for testing. Education stakeholders are exploring more efficient measurement tools in place of traditional paper-and-pencil tests (PPTs). Many of them foresee the promise of using computer-based testing (CBT) in their state assessment due to the advantages of CBTs over traditional PPTs in terms of immediate scoring and reporting of students' test results, greater test security, test administration efficiency, flexible test administration schedules, reduced costs compared to handling PPTs, the use of multimedia innovative item types that are not feasible in the PPT format, audio and large-print accommodations for vision-impaired students, and the ability to measure response time (Bennett, 2001, 2002; Boo & Vispoel, 1998; Folk & Smith, 1998; Klein & Hamilton, 1999; Parshall, Spray, Kalohn, & Davey, 2002; Schmit & Ryan, 1993). CBT can be administered via computer in the offline setting, in network configurations, or on the Internet. The application of CBT in state assessments is justified by the widespread availability of computers in schools. In addition, computer-based assessments have become a part of an integrated plan to apply technology throughout the educational process at school district, state, and national levels (Bennett, 2001, 2002; National Association of State Boards of Education, 2001; National Center for Education Statistics, 2000; National Commission on Excellence in Education, 1983).

Although CBT has gained in popularity for K–12 assessment in recent years, it has been extensively investigated before in the areas of licensure and certification tests. Some professional standards and guidelines have been well established to guarantee the proper development and use of CBT. These professional guidelines and standards address the issues regarding using CBT, including *Guidelines for Computerized Adaptive Test Development and Use in Education* (American Council on Education, 1995), *Guidelines for Computer-Based Testing* (Association of Test Publishers, 2000), *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999), *Guidelines for Computer-Based Tests and Interpretations* (APA, 1986), and *International Guidelines on Computer-Based and Internet-Delivered Testing*

(International Test Commission [ITC], 2004). According to the ITC guidelines (2004), major issues related to the application of CBT include computer hardware and software technology, test materials and testing procedure quality, control of the test delivery, test-taker authentication, prior practice, security issues of testing materials, privacy, data protection, and confidentiality. The ITC guidelines also emphasize the importance of guaranteeing that test developers, publishers, and users have sufficient knowledge and competence to ensure the appropriateness of using CBT. Test developers and publishers should ensure that psychometric standards have been met and evidence of the equivalence between the CBT and PPT versions of a test has been provided.

Most important, both the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and the *Guidelines for Computer-Based Tests and Interpretations* (APA, 1986) emphasize the significance of score equivalence. The score equivalence between CBT and PPT is defined as follows (APA, 1986):

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

The *Guidelines for Computer-Based Tests and Interpretations* (APA, 1986) also emphasizes the importance of eliminating irrelevant influences on test scores such as computer anxiety and computer experience. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) suggests that evidence of score equivalence should be provided to support any assertion that scores using different items or testing materials, different testing procedures, or test forms administered in different formats are interchangeable.

The factors that lead to the administration mode difference between CBT and PPT vary depending on different research studies. Some factors related to the presentation of items such as computer interface, item layout, and graphics in CBT may result in differences in examinee performances between CBT and PPT. Mazzeo and Harvey (1988) noted that tests that required multiscreen, graphical, or complex displays resulted in mode effects. Some computer-linked factors such as screen size, font size, and resolution of graphics may change the nature of a task so dramatically that CBT and PPT may no longer measure the same construct (McKee & Levinson, 1990). In addition, Vispoel, Wang, de la Torre, Bleiler, and Dings (1992) indicated that CBTs with or without item review do not necessarily yield equivalent results with PPT. However, other studies (Vispoel et al., 1992; Wise & Plake, 1989) showed that the inability to review and revise test response had a significant negative effect on examinee performance. Mueller and Wasser (1977) also suggested that item review was an important test-taking strategy that had a positive effect on examinee performance.

Test administration mode effects have been extensively studied. During the past 25 years, more than 300 studies have studied the test mode effects on intelligence, aptitude, ability, vocational interest, personality, and achievement tests. However, findings related to the score equivalence between CBT and PPT were not conclusive. Some studies indicate that the CBT scores were equivalent to the PPT scores (Bergstrom, 1992; Boo & Vispoel, 1998; Bugbee, 1996; Chin & Donn, 1991; Choi & Tinkler, 2002; Evans, Tannehill, & Martin, 1995; Johnson & Green, 2004; Neuman & Baydoun, 1998; Wang, Newman, & Witt, 2000), whereas other studies indicated that the results from CBT and PPT could not be used interchangeably (Godwin, 1999; Mazzeo & Harvey, 1988; Mead & Drasgow, 1993; Pommerich & Burden, 2000).

From a test-taker's perspective (Park, 2003), computerized assessment was easier. The attitudes expressed by test takers were generally more positive toward CBT than PPT (Wang, Young, & Brooks, 2004). Recent research (Russell, 1999; Russell & Haney, 1997; Russell & Plati, 2001a, 2001b) suggested that some students were more comfortable and accustomed to writing via the computer and that CBT may be a better option than PPT to assess students' writing ability.

Four previous meta-analysis studies (Bergstrom, 1992; Kim, 1999; Mead & Drasgow, 1993; Wang, Jiao, Young, Brooks, & Olson, 2007) examined the CBT and PPT mode effect on tests that measure general aptitude, ability, and achievement. Bergstrom (1992) compared the results of 20 comparability studies from eight research reports; 12 studies represented tests of adults, and the remaining 8 studies represented results for K–12 students. After removing 5 studies that contributed to the heterogeneity of effect size (ES) because they had the largest disproportional sample sizes (unbalanced sample size) between PPT and CBT, the results of the remaining 15 studies showed that the weighted mean ES between CBTs and PPTs was not statistically significant. However, by examining a moderator variable of mode order, the PPT had a higher mean score than the CBT when the examinee took both tests.

In the Mead and Drasgow (1993) study, comparability studies between CBT and PPT that measured young adults' and adults' cognitive ability were synthesized. Among 159 cross-mode correlations after correcting for measurement error, 123 were from timed power tests and 36 were from speeded tests. The overall corrected cross-mode correlation was .91, and the moderator variable of speededness had a moderate effect on administration mode. The computer delivery algorithm, that is, linear or adaptive computer tests, did not result in any differences between CBT and PPT scores.

Kim (1999) synthesized 226 ESs from 51 primary studies that included ability measures given as both PPT and either linear CBT or adaptive CBT. Among these studies, 4% of the samples were from K–12 students and 96% of the samples were from other educational institutes. This study reported, on average, that CBT and PPT were equivalent and CBT seemed easier than PPT for high school students.

The results also showed that the type of CBT (linear or adaptive) was the most important variable when evaluating the equivalence between CBT and PPT and that the equivalence between CBT and PPT held for mathematics or other cognitive measures but failed for English tests and other subject tests (science, medical knowledge, mechanical knowledge, education, etc.).

Goldberg, Russell, and Cook (2003) conducted a meta-analysis of mode effects between CBT and PPT specifically on K–12 writing assessments. They synthesized 26 studies conducted from 1992 to 2002 and found that mean ES was significantly higher for CBT than for PPT for quantity of writing and quality of writing. Their results showed that students who used computers when learning to write were not only more engaged and motivated in their writing, but also produced higher quality written work. However, this meta-analysis focused more on instruction than on assessment.

Wang et al. (2007) conducted a meta-analysis of CBT and PPT administration mode effects on K–12 student mathematics tests. Both initial and final results based on fixed- and random-effects models were presented. The results based on the final selected studies with homogeneous ESs showed that the administration mode had no statistically significant effect on K–12 student mathematics tests. Only the variable of computer delivery algorithm contributed to predicting the ES. The differences in scores between test modes were larger for linear tests than for adaptive tests. However, such variables as study design, grade level, sample size, type of test, computer delivery method, and computer practice did not lead to differences in student mathematics scores between CBT and PPT modes.

In addition to these meta-analyses, Mazzeo and Harvey (1988) conducted a review of the literature about the equivalence of scores from automated and conventional educational and psychological tests. However, limited ESs of CBT and PPT were reported only for two particular psychological tests.

Although previous studies provide insights into the effects of test administration mode on different achievement and ability measures of K–12, postsecondary, and adult learners, none of them specifically focused on K–12 students and their reading achievement and ability. The unique features of computerized reading tests—such as built-in features of CBT related to scrolling text, going back and forth to items of a particular passage in a testlet form, highlighting a passage, dealing with pop-up notes, zooming, and so forth—might cause difference between test modes. Given the fact that the findings from previous CBT and PPT comparability studies were not consistent and the focus of the previous meta-analysis of administration mode effects between CBT and PPT were not on K–12 students' reading assessment, it is necessary to synthesize the results from comparability studies that focus on K–12 students' reading assessments.

In addition, given the fact that more and more tests are now being administered because of the requirements of the NCLB Act on state assessment programs, it is crucial to understand better the impact of different test administration modes on the

scores. Furthermore, there is an increasing need from test developers', test publishers', and test users' points of view to know the direction and magnitude of the effects of the computer mode on K–12 students' reading achievement and ability across studies.

The purpose of this study is to synthesize the impact of administration mode on K–12 student reading tests. This study is a companion piece to Wang et al. (2007) and closely follows the aims and methodology of this earlier work. The current study focuses on K–12 students and the comparability of their test scores from CBT and PPT reading tests.

Method

Sample

Literature searches were manually and electronically conducted related to both published and unpublished studies of the CBT and PPT administration effects on students' test results. To avoid the biased retrieval of searching only major journals and readily retrievable studies, an exhaustive list of articles was selected by means of multiple procedures. The major sources of the literature search included the following:

- Journals (both e-journals/journals on CDs and manual searches, 1980–2005): *Applied Measurement in Education*, *Applied Psychological Measurement*, *Journal of Educational Measurement*, *Educational and Psychological Measurement*, *Psychological Methods*, *Psychological Bulletin*, *Journal of Technology, Learning, and Assessment*, *Computers in Human Behavior*, and *Computers & Education*.
- Databases (1980–2004): ERIC, Academic Search Elite, Expanded Academic ASAP, Ingenta, PsycINFO, Dissertation Abstracts, ProQuest, and Ovid.
- Test publisher Web sites: ETS, ACT, CTB, Harcourt, Pearson, Prometric, Riverside Publishing, University of Iowa's ITBS, departments of education (state and federal), and Web search engines (Google and Yahoo) with the keywords "computer-based test," "computerized test," "computer-based and paper-and-pencil tests," "administration mode effect," "equivalence study," "comparability study," "mode effect on students reading tests," and others.
- Manual searches in a university library.
- Personal contacts.

The initial literature search resulted in 312 articles.

Criteria for Study Inclusion

A study was included in this review only if it was possible to calculate an ES estimate of the difference in reading scores between the CBT and PPT for K–12 students. Because multiple results may be reported from the same study, the choice had to be made to see if these results could be treated as if they were from separate

independent studies. After carefully examining each of the studies, it was found that almost all multiple results reported for the same study used student samples from different grades. Because there was no dependence of scores within a single outcome measure for each result, it is reasonable to treat multiple results as though they were from separate independent studies. Each study had to meet the following inclusion criteria:

- The study had to be conducted between 1980 and 2005.
- The samples of study had to be drawn from the K–12 student population, and the within-group sample size had to be larger than 25.
- The study should have quantitative outcome measures (mean and standard deviation) of one of student achievement, aptitude, or ability of reading on both CBT and PPT.
- The study should have the design to compare the scores from both CBT and PPT.
- The test language in the study must be English because the major target population (U.S. K–12 students) in this report uses English.

These criteria yielded a sample of 11 primary studies that contained 42 independent experiments or data sets. A descriptive summary of the individual studies included in this review is presented in Table 1. The selected studies are marked with an asterisk in the references.

Moderators

The following four category attributes were coded to describe each study:

1. Attributes of the article: publication name, year of publication, and publication type.
2. Attributes of the test: test name, type of test, test content, test length, computer delivery method, and computer delivery algorithm.
3. Characteristics of the study: design of study, test mode balance, outcome measure, and computer practice availability.
4. Attributes of the sample: sample size, target population or grade, and whether information was included about gender, ethnicity, school setting, and students' socioeconomic status.

Each study was reviewed by two researchers. The potential moderator variables were coded independently by the two researchers. Intercoder agreement ranged from a low of 89% for the characteristics of the study to a high of 99% for the attributes of the article. Disagreements were resolved through discussion.

Issues of ES Estimation Procedures

In this study, 22 out of 42 studies (more than half) used repeated-measures designs in which the same students were tested under both CBT and PPT conditions

Table 1
The Summary of Studies ($N = 42$)

Author (Publication Year)	CBT Sample Size	PPT Sample Size	Sample Grade	Test Type	Outcome Measure	Design	Mode Order Balanced	CBT Delivery Algorithm
Arce-Ferrer et al. (2004)	1,511	1,511	5	State achievement test	Reading total	Nonrandom	Yes	Linear
Eignor (1993)	271	271	9–12	National ability test	Reading total	Random	Yes	CAT
Eignor (1993)	271	271	9–12	National ability test	Reading total	Random	Yes	CAT
Ito et al. (2004)	2,239	717	4, 5	National ability test	Verbal words total	Random	Yes	Linear
Ito et al. (2004)	2,202	716	4, 5	National ability test	Verbal context total	Random	Yes	Linear
Ito et al. (2004)	1,781	570	6, 7	National ability test	Verbal words total	Random	Yes	Linear
Ito et al. (2004)	1,736	564	6, 7	National ability test	Verbal context total	Random	Yes	Linear
Ito et al. (2004)	1,407	302	8, 9	National ability test	Verbal words total	Random	Yes	Linear
Ito et al. (2004)	1,373	303	8, 9	National ability test	Verbal context total	Random	Yes	Linear
Ito et al. (2004)	528	225	10, 11	National ability test	Verbal words total	Random	Yes	Linear
Ito et al. (2004)	518	225	10, 11	National ability test	Verbal context total	Random	Yes	Linear
Ito et al. (2004)	621	183	11, 12	National ability test	Verbal words total	Random	Yes	Linear
Ito et al. (2004)	623	185	11, 12	National ability test	Verbal context total	Random	Yes	Linear
Kingsbury et al. (1988)	1,331	1,678	3–8	District achievement test	Reading total	Nonrandom	No	CAT
Kingsbury (2002)	1,331	1,678	4	State achievement test	Reading total	Nonrandom	No	CAT
Kingsbury (2002)	443	443	5	State achievement test	Reading total	Nonrandom	Yes	CAT
Poggio et al. (2005)	3,228	3,228	5	State achievement test	Reading total	Nonrandom	No	Linear
Poggio et al. (2005)	2,751	2,751	8	State achievement test	Reading total	Nonrandom	No	Linear
Poggio et al. (2005)	784	784	11	State achievement test	Reading total	Nonrandom	No	Linear
Poggio et al. (2005)	385	385	5	State achievement test	Reading total	Nonrandom	Yes	Linear
Poggio et al. (2005)	370	370	8	State achievement test	Reading total	Nonrandom	Yes	Linear
Pommerich (2004)	908	985	11, 12	Researcher-developed test	Reading total	Random	Yes	Linear
Pommerich (2004)	996	1,086	11, 12	Researcher-developed test	Reading total	Random	Yes	Linear
Pommerich (2004)	1,089	1,086	11, 12	Researcher-developed test	Reading total	Random	Yes	Linear
Pomplun et al. (2002)	94	94	10–12	National aptitude test	Reading total	Random	Yes	Linear

Pomplun et al. (2002)	94	94	10–12	National aptitude test	Reading total	Random	Yes	Linear
Pomplun et al. (2000)	84	84	11–12	National aptitude test	Reading total	Nonrandom	Yes	Linear
Pomplun et al. (2000)	101	101	11–12	National aptitude test	Reading total	Nonrandom	Yes	Linear
Schwartz et al. (2003)	207	1,522	4–9	National aptitude test	Reading total	Nonrandom	No	Linear
Schwartz et al. (2003)	219	1,690	4–9	National aptitude test	Reading total	Random	No	Linear
Wang et al. (2004)	120	120	2	National diagnostics test	Reading total	Random	Yes	Linear
Wang et al. (2004)	205	205	3	National diagnostics test	Reading total	Random	Yes	Linear
Wang et al. (2004)	231	231	4	National diagnostics test	Reading total	Random	Yes	Linear
Wang et al. (2004)	267	267	5	National diagnostics test	Reading total	Random	Yes	Linear
Wang et al. (2004)	283	283	6	National diagnostics test	Reading total	Random	Yes	Linear
Wang et al. (2004)	297	297	7	National diagnostics test	Reading total	Random	Yes	Linear
Wang et al. (2004)	318	318	8	National diagnostics test	Reading total	Random	Yes	Linear
Wang et al. (2004)	165	165	9	National diagnostics test	Reading total	Random	Yes	Linear
Wang et al. (2004)	176	176	9	National diagnostics test	Reading total	Random	Yes	Linear
Wang et al. (2004)	279	279	10	National diagnostics test	Reading total	Random	Yes	Linear
Wang et al. (2004)	263	263	11	National diagnostics test	Reading total	Random	Yes	Linear
Wang et al. (2004)	267	267	12	National diagnostics test	Reading total	Random	Yes	Linear

Note: CAT = computerized adaptive testing; CBT = computer-based test; PPT = paper-and-pencil test; FD = different forms administrated for each mode; FS = same forms administrated for both modes.

with either counterbalancing or random assignments to either of the two modes. The issue of whether ES across independent-groups (between-subjects) designs and repeated-measures (within-subjects) designs are comparable should be carefully considered. Previous studies (Dunlop, Cortina, Vaslow, & Burke, 1996; Morris & DeShon, 2002) suggested the ESs from two designs should not be combined unless the researcher can justify doing so based on rational analysis. To estimate ES for the repeated-measures design, both Becker (1988) and Dunlop et al.'s equations need correlation between the experiment and control groups. However, most of these studies that used the repeated-measures design did not present the correlation information. Therefore, in this study, the ES must be estimated from the mean and standard deviation. The consequence of using test statistics that do not take the correlation between the measures into account is the overestimation or inflation of the ES. Therefore, the overall result of the calculated ESs that ignore the distinction between these two types of designs can be regarded as the upper bound of the actual ESs.

Data Analysis

Because the purpose of this study is to examine the difference of test scores across modes, the *d*-type ES is preferred over the *r*-type ES. The standardized mean difference, as ES represents a standardized group (CBT vs. PPT), contrasts on an inherently continuous measure (reading score). The ES is calculated based on Hedges's (1981, 1987; Hedges & Olkin, 1985) *g* formulations of meta-analysis, which ignores the difference between research designs completely (Dunlop et al., 1996). It is the mean difference *d* between the CBT and PPT test scores, divided by the pooled standard deviation. A positive ES indicates that CBT has a higher score than PPT on the reading test in question. Regardless of the research design used, the *g* statistics overestimate the population ES, especially for smaller sample sizes, so the *g* is converted to *d* to correct bias (Hedges & Olkin, 1985).

The standardized mean difference or ES *g* for any individual study is defined (Hedges, 1981; Hedges & Olkin, 1985) as follows:

$$g = \frac{M_{CBT} - M_{PPT}}{SD_p}, \tag{1}$$

where M_{CBT} and M_{PPT} are the means of CBT and PPT scores and SD_p is the pooled estimate of the standard deviation,

$$SD_p = \sqrt{\frac{(n_{CBT} - 1)sd_{CBT}^2 + (n_{PPT} - 1)sd_{PPT}^2}{n_{CBT} + n_{PPT} - 2}}, \tag{2}$$

where n_{CBT} and n_{PPT} are the sample sizes of CBT and PPT and sd_{CBT} and sd_{PPT} are the standard deviations, respectively. The transformation of *g* to *d* (unbiased ES) corrects for small sample size bias:

$$d = g\left(1 - \frac{3}{4N - 9}\right), \tag{3}$$

with the estimate of sampling error variance of the d statistics if sample sizes are quite unequal (Hedges & Olkin, 1985):

$$\mathrm{var}(d) = \frac{n_{CBT} + n_{PPT}}{n_{CBT} n_{PPT}} + \frac{d^2}{2(n_{CBT} + n_{PPT})}. \tag{4}$$

Because each of the independent studies shares a common d of ES and studies vary in sample size, the d estimated based on a large sample size is more precise than the d from a small sample size. Common practice is to give the study with a large sample size more weight than the study with a small sample size. One of the weighting approaches is to weigh estimators by giving weight inversely proportional to the variance in each study (Hedges & Olkin, 1985). The weight w_i for study i gives optimal weight that minimizes the variance of d_i:

$$w_i = \frac{1}{\mathrm{var}(d_i)}. \tag{5}$$

Then, d_w of weighted mean ES can be expressed as follows:

$$d_w = \frac{\sum_i d_i w_i}{\sum_i w_i} = \frac{\sum_i d_i / \mathrm{var}_i}{\sum_i 1/\mathrm{var}_i}. \tag{6}$$

Statistical significance of the mean ES is assessed by calculating the 95% confidence interval (CI) for the population parameter. A significance level of .05 is used to indicate the significance when zero is not within the 95% CI.

The homogeneity of ES needs to be evaluated before the final conclusion can be drawn. The consistency of the results in different studies can be analyzed by the homogeneity test using Q statistics:

$$Q = \sum_i \frac{(d_i - d_w)^2}{\mathrm{var}_i}. \tag{7}$$

The large-sample statistic Q is approximately distributed as a χ^2 distribution with degrees of freedom = number of ds $- 1$. If the null hypothesis that all ds are equal is rejected, the estimated ds should not be pooled because they do not estimate the same parameter. Then those studies that cause the heterogeneity of the selected studies are carefully examined and excluded from the analysis. The homogeneity test is rerun using Q statistics until the remaining studies have homogeneous ESs. Once the remaining studies have homogeneous ESs, the significance test of the mean ES between CBT and PPT is rerun. The conclusion is drawn based on the homogeneous ESs.

Results

Sample of Studies and Data Sets

The characteristics of the selected studies included in this meta-analysis study are summarized in Table 2. Most studies were conducted in 2004, which reflected the current trend of increase in using CBT. The majority of the studies were conference presentations and research or technical reports from testing companies or research organizations. More than 80% of the samples used in these studies were from middle schools and high schools. More than 70% of the studies used an experimental design with randomization. Test mode order was considered in about 90% of the studies. More than half of the studies used a sample size larger than 400. The other half used a sample size from 100 to 400. Three types of tests—national achievement tests, national aptitude/ability/diagnostic tests, and state-specific tests—encompassed the majority of the studies. About 83% of the CBTs used the fixed linear delivery algorithm. Only about 16% of the included studies used the computerized adaptive testing algorithm. More than 45% of the CBTs were administered on individual PCs. Others were administered either on the Internet or on the local network/Web. About 90% of the included studies reported no PC experience of students who participated in the comparison study. The other samples had previous PC experience.

Weighted ESs and Homogeneity Analysis

The summary of the study ESs (Hedges g and unbiased d) given in Table 3 allows a determination of whether students' reading test scores differed by using different administration modes. Based on Equation 1, a negative ES (g or d) indicates that CBT has a lower score than PPT; a positive g or d indicates that CBT has a higher score than PPT. Among the 42 studies, the 95% CIs for 12 (28.6%) of the studies did not contain zero, which means that the differences between CBT and PPT were statistically significant. The fixed-effects estimate of the overall weighted mean d_w was $-.077$ with a 95% CI of $[-.094, -.060]$; thus, this estimate was statistically significant at the α level of .01 ($p = .000$). This indicates that examinees scored significantly higher on PPT than on CBT. Because of the diverse characteristics of the studies included in this analysis, the fixed-effects model showed that the homogeneity of ESs for the 42 studies was rejected, $Q(41) = 356.54$, $df = 41$, $p < .01$. Although ESs are statistically significant, all of them can be classified as practicably negligible per Cohen (1988) criteria on ES (less than .2 is negligible, .2 is small, .5 is medium, and .8 is large).

If the distribution of ESs is assumed to be heterogeneous based on the fixed-effects model, then a random-effects model is fitted to the data. The random-effects model assumes that the variability among ESs is the combination of both sampling errors and differences in true population ESs (Lipsey & Wilson, 2001). After fitting

Table 2
Descriptive Summary of Study Attributes

Attributes	Value
Study characteristics	
Mean year of study (range = 1988–2005, *SD* = 3.46 years)	2002.79
Median year of study	2004
Mode year of study (70%)	2004
Publication type (%)	
Refereed journal paper	18.9
Conference presentation	45.9
Research/technical report	35.1
Sample grade (%)	
Elementary school	24.3
Middle school	18.9
High school	48.6
Others	8.1
Characteristics of moderators	
Design (%)	
Quasi-experimental/nonrandomized	28.6
Experimental/randomized	71.4
Sample size (%)	
$50 \leq N < 100$	7.1
$100 \leq N < 200$	9.5
$200 \leq N < 400$	31.0
$400 \leq N$	52.4
Test type (%)	
National achievement test	28.6
National aptitude/ability/diagnostics test	45.2
State-specific test	19.0
District/school-specific test	2.4
Unspecified	4.8
Test mode order considered (%)	
Yes	89.2
No	10.8
Computer delivery algorithm (%)	
Linear (fixed)	83.3
Computerized adaptive testing	16.7
Computer administration method (%)	
Individual PC	47.6
Local network/Web	7.2
Internet	35.7
Unspecified	9.5
Outcome measure type (%)	
Raw score	88.1
Scale score	11.9
PC experience (%)	
Yes	9.5
No	90.5

Table 3
The Effect Sizes (Random Effects) of All Studies ($N = 42$, $Q = 356.54$, $df = 41$, $p < .01$)

Author (Publication Year)	Study ID	Sample Grade	Pooled *SD*	Hedges *g*	Unbiased *d*	95% CI *
Arce-Ferrer et al. (2004)	1	5	8.029	0.102	0.102	0.17/0.03 *
Eignor (1993)	2	9–12	17.820	0.053	0.053	0.22/−0.12
Eignor (1993)	3	9–12	17.367	0.014	0.014	0.18/−0.15
Ito et al. (2004)	4[a]	4, 5	3.861	−0.308	−0.308	−0.22/−0.39 *
Ito et al. (2004)	5[a]	4, 5	4.166	−0.562	−0.562	−0.48/−0.65 *
Ito et al. (2004)	6[a]	6, 7	4.027	−0.308	−0.308	−0.21/−0.40 *
Ito et al. (2004)	7[a]	6, 7	3.991	−0.418	−0.418	−0.32/−0.51 *
Ito et al. (2004)	8[a]	8, 9	4.248	−0.388	−0.388	−0.26/−0.51 *
Ito et al. (2004)	9[a]	8, 9	4.022	−0.211	−0.211	−0.09/−0.34 *
Ito et al. (2004)	10	10, 11	4.019	−0.002	−0.002	0.15/−0.16
Ito et al. (2004)	11	10, 11	4.049	−0.151	−0.150	0.01/−0.31
Ito et al. (2004)	12	11, 12	4.358	−0.124	−0.124	0.04/−0.29
Ito et al. (2004)	13	11, 12	4.136	−0.206	−0.205	−0.04/−0.37 *
Kingsbury et al. (1988)	14	3–8	15.773	0.174	0.174	0.31/0.04 *
Kingsbury (2002)	15	4	12.200	0.007	0.007	0.08/−0.07
Kingsbury (2002)	16	5	12.538	−0.018	−0.018	0.05/−0.09
Poggio et al. (2005)	17	5	10.072	−0.061	−0.061	−0.01/−0.11 *
Poggio et al. (2005)	18	8	8.752	0.001	0.001	0.05/−0.05
Poggio et al. (2005)	19	11	9.906	−0.018	−0.018	0.08/−0.12
Poggio et al. (2005)	20	5	9.046	0.048	0.047	0.19/−0.09
Poggio et al. (2005)	21	8	9.022	0.154	0.154	0.30/0.01 *
Pommerich (2004)	22	11, 12	7.254	−0.145	−0.145	−0.05/−0.23 *
Pommerich (2004)	23	11, 12	7.120	−0.029	−0.029	0.06/−0.12
Pommerich (2004)	24	11, 12	7.011	−0.036	−0.036	0.05/−0.12
Pomplun et al. (2002)	25	10–12	27.380	0.064	0.063	0.35/−0.22
Pomplun et al. (2002)	26	10–12	28.880	0.045	0.044	0.33/−0.24
Pomplun et al. (2000)	27	11–12	24.808	0.019	0.019	0.32/−0.28
Pomplun et al. (2000)	28	11–12	37.139	−0.116	−0.115	0.16/−0.39
Schwartz et al. (2003)	29	4–9	10.343	−0.058	−0.058	0.09/−0.20
Schwartz et al. (2003)	30	4–9	9.823	0.112	0.112	0.25/−0.03
Wang et al. (2004)	31	2	15.600	−0.152	−0.151	0.10/−0.40
Wang et al. (2004)	32	3	16.697	0.018	0.018	0.21/−0.18
Wang et al. (2004)	33	4	16.173	−0.030	−0.030	0.15/−0.21
Wang et al. (2004)	34	5	13.573	0.029	0.029	0.20/−0.14
Wang et al. (2004)	35	6	13.277	0.035	0.035	0.20/−0.13
Wang et al. (2004)	36	7	14.651	0.029	0.029	0.19/−0.13
Wang et al. (2004)	37	8	13.492	0.001	0.001	0.16/−0.15
Wang et al. (2004)	38	9	11.114	0.065	0.065	0.28/−0.15
Wang et al. (2004)	39	9	13.891	0.022	0.022	0.23/−0.19
Wang et al. (2004)	40	10	13.554	0.037	0.037	0.20/−0.13

Table 3 (continued)

Author (Publication Year)	Study ID	Sample Grade	Pooled *SD*	Hedges *g*	Unbiased *d*	95% CI *
Wang et al. (2004)	41	11	13.212	0.040	0.040	0.21/−0.13
Wang et al. (2004)	42	12	12.317	0.009	0.009	0.18/−0.16

Note: Negative effect sizes (*g*s or *d*s) indicate that the CBT had a lower score than the PPT, and positive *g*s or *d*s indicate that the CBT had a higher score than the PPT. CI = confidence interval; Q = homogeneity of *d*s.

a. Indicates the excluded study.

* Represents statistically significant effect size.

a random-effects model, the test of homogeneity of ESs and outlier analysis were conducted to evaluate whether the overall ES was homogeneous when the most deviant study outcomes were excluded from computing the mean effect. After removing 6 studies that had the largest differences of sample sizes between CBT and PPT, the null hypothesis of the homogeneity of the ESs for the selected 36 studies was no longer rejected at the α level of .01: $Q(35) = 54.75$ and $p = .018$. The random-effects estimate of the overall weighted mean d_w for the full sample was −.060 with a 95% CI of [−.114,−.006] and was statistically significant at the α level of .01 ($p = .028$). The d_w based on the selected sample, however, was −.004 with a 95% CI of [−.031, .023] and was no longer statistically significant at the α level of .01 ($p = .782$), which indicates that examinees' performance on PPT was not significantly better than their performance on CBT.

Besides traditional statistics concerns, the rationale to remove six samples from our study is that for these six samples, CBT consistently displayed a pattern of steadily increasing percentages of unrecorded responses because of speededness. The author reported that between 9% and 17% of students responded in the CBT survey that they did not have enough time to respond to all the test items. This phenomenon is really unusual for a general standardized achievement test. Although there are time limits, most achievement tests are power tests, not speeded tests.

Moderator Analysis

Because the statistically significant variability or heterogeneous distributions in the ESs across studies were not due to random error but to moderator variables, it is important to determine the condition under which the ESs may vary. The weighted multiple regression analysis with the random-effects model can be used to evaluate moderator variables (Lipsey & Wilson, 2001). The effect of individual moderators on the ES is investigated. The statistically significant moderators include study design, sample size, computer practice provided (all three with negative coefficients), and computer delivery algorithm (with a positive coefficient). The rest of the moderators

Table 4
The Results of Meta-Analysis Modified Weighted Multiple Regression

Variable	*B*	*SE*	−95% CI	95% CI	*Z*	*p* value	Beta
Constant	1.06	0.18	0.71	1.42	5.86	.00	.00
Study design	−0.16	0.04	−0.23	−0.08	−4.15	.00	−.49
Grade level	−0.01	0.01	−0.03	0.01	−1.30	.19	−.10
Sample size	−0.13	0.03	−0.18	−0.07	−4.70	.00	−.44
Type of test	−0.04	0.02	−0.07	0.00	−1.90	.06	−.19
Computer delivery method	0.01	0.01	−0.02	0.03	0.45	.66	.05
Computer delivery algorithm	0.05	0.01	0.02	0.07	3.39	.00	.22
Computer practice provide	−0.07	0.03	−0.12	−0.02	−2.89	.00	−.24

(grade level, type of test, and computer delivery method) were not statistically significant, and this suggests that they were not statistically significant in contributing to predicting the ES.

The section related to the homogeneity analysis in Table 4 presents the Q statistics for the model and the Q statistics for the residual. The former indicates if the regression model explains a significant portion of the variability across ESs, and the latter indicates if the remaining variability across ESs is homogeneous. For this study, the regression model explained a significant portion of the variability across the ESs ($p < .01$), and the remaining variability across ESs was homogeneous. In total, the 10 moderators (predictors) accounted for 33.3% of the total variance in the dependent variable, which was the ES. The effect of individual moderators on ES was also provided. The study design, sample size, and computer practice provided are significant moderators with negative coefficients, which means that providing information or not systematically affected the ES. The negative Beta indicates that studies providing such information tended to have greater ESs than those not providing information. Another significant moderator is the computer delivery algorithm with a positive coefficient, which means that the mean ES of the linear fixed-form algorithm is relatively higher than those of the computerized adaptive test algorithm. Among the rest of the moderators (study design, sample size, type of test, computer delivery method, computer practice provided, publication type, and ethnicity), none were statistically significant.

Summary

This meta-analysis study specifically focused on the effect of test administration mode on K–12 students' reading achievement and ability tests. The initial results based on the 42 ESs using both fixed-effects and random-effects models showed that the Q statistics of the homogeneity test rejected the null hypothesis of homogeneity of ESs and that the estimate of overall weighted mean ES d_w was statistically

significant. After removing 6 studies based on outlier analysis, the null hypothesis of homogeneity of ESs for the remaining 36 studies was no longer rejected. The variability across the ESs did not exceed what would be expected from sampling errors based on the random-effects model. The conclusion from the meta-analyses of the 36 homogeneous studies conducted in the last 20 years was that the difference between students' reading achievement scores from CBT- and PPT-administrated tests was not statistically significant. This result is consistent with the findings from our previous study on mathematics (Wang et al., 2007). The statistically significant moderators include study design, sample size, and computer practice provided with negative coefficients, and computer delivery algorithm with a positive coefficient. The rest of the moderators (grade level, type of test, and computer delivery method) were not statistically significant, and this suggests that they were not statistically significant in contributing to predicting the ES.

Kolen and Brennan (1995) suggested that factors that affected the validity of CBTs are likely test specific. Therefore, analyses for mode effect are necessary for any test offered in both modes. Green, Bock, Humphreys, Linn, and Reckase (1984) believed that CBT and PPT are equally valid only if they have been demonstrated to yield equivalent measures. Other researchers reaffirmed the need to determine empirically the equivalence of CBT and PPT. To achieve score equivalence between CBT and PPT versions of the same test, all studies agreed that the forms must be scrutinized to the same degree as would be required for the construction of parallel test forms. When a CBT has been designed to be used and interpreted as a parallel version to an existing PPT, equivalences in terms of item characteristics, scores, test constructs, and examinee behavior between the two versions of the test can be used as evidence of the validity for CBT (Green et al., 1984; Parshall et al., 2002). In addition to the factors mentioned above, the effect of mode on passing rate for a given achievement level and score distribution should also be investigated. In practice, it will put a lot of burden on test practitioners and test users to verify the equivalence between CBT and PPT every time they create a new CBT form. Because meta-analysis can generalize the results from multiple comparability studies, the results from meta-analysis will provide an overall view of the equivalence between CBT and PPT. However, this does not mean that comparability studies are not needed in specific circumstances.

References

References marked with an asterisk indicate studies included in the meta-analysis.

American Council on Education. (1995). *Guidelines for computerized adaptive test development and use in education.* Washington, DC: Author.

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

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

* Arce-Ferrer, A., Lau, C. A., & Griph, G. (2004). *Comparison of paper-and-pencil and online versions of Grade 5 mathematics and Grade 8 reading and writing tests*. San Antonio, TX: Harcourt Assessment.

Association of Test Publishers. (2000). *Guidelines for computer-based testing*. Washington, DC: Author.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*, 257-278.

Bennett, R. E. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archive*, *9*(5). Retrieved June 20, 2004, from http://epaa.asu.edu/epaa/v9n5.html

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment*, *1*(1). Available at http://www.bc.edu/research/intasc/jtla/journal/pdf/v1n1_jtla.pdf

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, CA.

Boo, J., & Vispoel, W. P. (1998, April). *Computer versus paper-pencil assessment of educational development: Score comparability and examinee preference.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Bugbee, A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, *28*, 282-299.

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Dunlop, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures design. *Psychological Methods*, *1*, 170-177.

* Eignor, D. R. (1993). *Deriving comparable scores for computer adaptive and conventional tests: An example using the SAT* (Research Report RR-93-55). Princeton, NJ: Educational Testing Service.

Evans, L. D., Tannehill, R., & Martin, S. (1995). Children's reading skills: A comparison of traditional and computerized assessment. *Behavior Research Methods, Instruments, & Computers*, *27*(2), 162-165.

Folk, V. G., & Smith, R. (1998, September). *Model for delivery of computer-based tests.* Paper presented at the ETS-sponsored colloquium on Computer-Based Tests: Building the Foundation for Future Assessments, Philadelphia, PA.

Godwin, J. (1999, April). *Designing the ACT ESL listening test.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *Journal of Technology, Learning, and Assessment*, *2*(1). Available at http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1007&context=jtla

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347-360.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107-128.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science: The empirical cumulativeness of research. *American Psychologist*, *42*, 443-455.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

International Test Commission. (2004). *International guidelines on computer-based and Internet-delivered testing*. Available from http://www.intestcom.org/itc_projects.htm

* Ito, K., & Sykes, R. C. (2004, April). *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.

Johnson, M., & Green, S. (2004). *On-line assessment: The impact of mode on student performance.* Paper presented at the British Educational Research Association Annual Conference, Manchester, UK.

Kim, J. (1999, October). *Meta-analysis of equivalence of computerized and P&P tests on ability measures.* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chicago, IL.

* Kingsbury, G. G. (2002, April). *An empirical comparison of achievement level estimates from adaptive tests and paper-and-pencil tests.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

* Kingsbury, G. G., & Houser, R. L. (1988, April). *A comparison of achievement level estimates from computerized adaptive testing and paper-and-pencil testing.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Klein, S. P., & Hamilton, L. (1999). *Large-scale testing: Current practices and new directions* (Research Report IP-182). Santa Monica, CA: RAND.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature* (College Board Rep. No. 88-8, ETS RR No. 88-21). Princeton, NJ: Educational Testing Service.

McKee, L. M., & Levinson, E. M. (1990). A review of the computerized version of the Self-Directed Search. *Career Development Quarterly*, *38*, 325-333.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *9*, 287-304.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105-125.

Mueller, D. J., & Wasser, V. (1977). Implications of changing answers on objective test items. *Journal of Educational Measurement*, *14*, 9-14.

National Association of State Boards of Education. (2001). *Any time, any place, any path, any pace: Taking the lead on e-learning policy.* Retrieved June 10, 2003, from http://www.nasbe.org/Organization_Information/e_learning.pdf

National Center for Education Statistics. (2000). *Teacher use of computers and the Internet in public schools* (NCES 2000-090). Retrieved June 10, 2003, from http://nces.ed.gov/pubs2000/2000090.pdf

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Retrieved June 10, 2003, from http://www.ed.gov/pubs/NatAtRisk

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

Park, J. (2003). A test-taker's perspective. *Education Week*, *22*(35), 15.

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

* Poggio, J., Glasnapp, D., Yang, X., Beauchamp, A., & Dunham, M. (2005, April). *Assessing reading comprehension based on computerized and paper and pencil formats in a state large scale assessment program.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

* 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 at http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1002&context=jtla

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, LA. Available at http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1002&context=jtla

* Pomplun, M., Frey, S., & Becker, D. (2000). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement*, *62*, 337-354.

* Pomplun, M., Frey, S., Becker, D., & Hughes, K. (2002, April). *The validity of a computerized measure of reading rate.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Russell, M. (1999). Testing writing on computers: A follow-up study comparing performance on computer and paper. *Educational 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. *Educational Policy Analysis Archives*, *5*(3). Available from http://olam.ed.asu.edu/epaa/v5n3.html

Russell, M., & Plati, T. (2001a). Effects of computer versus paper administration of a state-mandated writing assessment. *Teachers College Record.* Available from http://www.tcrecord.org/PrintContent.asp?ContentID = 10709

Russell, M., & Plati, T. (2001b). Mode of administration effects on MCAS composition performance for Grade eight and ten. *Teachers College Record.* Available from http://www.tcrecord.org/Content.asp?ContentID = 10709

Schmit, M. J., & Ryan, A. M. (1993). Test-taking disposition: A missing link? *Journal of Applied Psychology*, *77*, 624-637.

* Schwarz, R. D., Rich, C., & Podrabsky, T. (2003). *A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Stenner, A. J. (1996, October). *Measuring reading comprehension with the Lexile framework*. Paper presented at the California Comparability Symposium, Burlingame, CA.

Vispoel, W. P., Wang, T., de la Torre, R., Bleiler, T., & Dings, J. (1992, April). *How review options and administration mode influence scores on computerized vocabulary tests.* Paper presented at the annual meeting of the AERA, San Francisco, CA. (ERIC Document Reproduction Service No. ED 346 161)

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

Wang, S., Newman, L, & Witt, E. A. (2000). *AT&T aptitude test equivalence study: A comparison of computer and paper-and-pencil employment examinations*. Bala Cynwyd, PA: Harcourt Assessment System.

* Wang, S., Young, M. J., & Brooks, T. E. (2004). *Administration mode comparability study for Stanford Diagnostic Reading and Mathematics Tests.* San Antonio, TX: Harcourt Assessment.

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.