What do we know about choosing to take a high-stakes test on a computer?

Florida Department of Education
Assessment and School Performance
(850) 245-0513

Prepared by CTB/McGraw-Hill under contract to the Florida Department of Education
September 4, 2006

What do we know about choosing to take a high-stakes test on a computer?

Introduction

The State of Florida, in general, requires students to pass both the Sunshine State Standards (SSS) Reading and the SSS Mathematics Grade 10 FCAT in order to receive a standard high school diploma. Florida provides students the opportunity to repeat the test either on a computer or by traditional paper-and-pencil format. To support schools, parents, and students in making the best possible choices of administration mode, the Department of Education is offering this paper that describes research findings regarding the comparability of test scores derived from computer-administered and paper-administered tests. We include references to the articles which are summarized here so that interested readers can consult the original sources.

Choosing between computer-administered and paper-administered tests would be easier if there were clear, incontrovertible evidence that for all students there is no difference in results whether a test is taken on computer or by printed test materials. Unfortunately, while the preponderance of the evidence suggests that for multiple-choice tests (such as Grade 10 FCAT) student performance does not *significantly* differ with regard to mode of administration, studies suggest that some students may do better on computer and other students may do better on paper. Also, in determining the practical significance of these studies, one must consider the stakes of the test, and there is no consensus as to how much difference based on mode of administration is too much.

To explain at least some of this inconsistency, there are two likely reasons:

First, not all computerized test administration systems are the same. Software developers have learned much over the years about how to make test administration software more user-friendly. For instance, some older systems required students to scroll both up and down and side to side to read an entire passage or large item. New systems recognize that this makes a test harder, and thus scrolling is minimized to a single dimension at most.

Second, the strongest (most informative) type of research study is one where students are randomly assigned to testing conditions regardless of test mode. However, it is often impractical to have half the students in a classroom testing on computers while the other half are testing on paper. Sometimes it is easier to let students (or teachers) determine the test mode. Results from this latter kind of study might be affected by who has chosen to participate in each group (for example, do students who have more familiarity with computers naturally choose to participate in the online group?).

Early comparability studies

Large-scale, computer-based testing has been around since the early 1980s. (Small systems for research purposes were used for at least a decade or two before that.) Early systems were primitive compared to current ones. Moreover, most early computer-administered testing programs were used by young adults in post-secondary educational settings, by professional adults seeking licensure or certification, or by those in the military. This paper focuses on studies published in the last ten years for which multiple-choice[1] educational achievement tests were administered to students in Grades K–12.

Comparability for K–12 students on multiple-choice tests

Paek (2005) presented a table summarizing the results of recent comparability studies. Table 1 is adapted from her work and updated to include additional studies. Adding up the number of grade/subject combinations, we see that out of 97 cases the results for 74 were deemed comparable, in 8 the computer-administered test appeared more difficult, and in 15 the paper test seemed more difficult.

Table 1
Recent K–12 Comparability Studies of Multiple-Choice Tests

	More Difficult Administration Mode		
	Computer	Paper	Comparable
Math	Choi & Tinkler (2002), G3 Cerillo & Davis (2004), Algebra Sandene, Bennett, Braswell, & Oranje (2005) Way, Davis, & Fitzpatrick (2006), G11	Choi & Tinkler (2002), G10	Kim & Hunyh (2006), Algebra Kingston (2002), G1,4,6,8 Pearson Educational Measurement [PEM] (2002), Algebra PEM (2003), Algebra II Nichols and Kirkpatrick (2005) Poggio, Glassnapp, Yang, & Poggio (2005), G7 Russell (1999), G8 Russell & Haney (1997), G6,7,8 Wang (2004), G2–5,7–12 Way, Davis, & Fitzpatrick (2006), G8
Language Arts		Russell & Haney (1997), G6,7,8	Kim & Huynh (2006), HS Kingston (2002), G1,4,6,8 Pommerich (2004), G11–12 Russell (1999), G8 Russell & Haney (1997, 2002), G6,7,8 Way, Davis, & Fitzpatrick (2006), G11

[1] Comparability of computer-administered and paper-and-pencil constructed-response tests is a more complex issue and is outside the scope of this paper.

Table 1 (continued)
Recent K–12 Comparability Studies of Multiple-Choice Tests

	More Difficult Administration Mode		
	Computer	Paper	Comparable
Reading	Choi & Tinkler (2002), G3 Cerillo & Davis (2004), HS English Way, Davis, & Fitzpatrick (2006), G8	Choi & Tinkler (2002), G10 Pomplun, Frey, & Becker (2002), HS O'Malley, et al. (2005), G2–5,8	Kingston (2002), G1,4,6,8 Nichols & Kirkpatrick (2005), PEM (2002), HS English Pommerich (2004), G11,12 Russell (1999), G8 Russell & Haney (1997), G6,7,8 Wang (2004), G2–5,7–12
Science	Cerillo & Davis (2004), Biology	Russell (1999), G8 Russell & Haney (1997, 2002), G6,7,8	Kim & Huynh (2006), Physical Science, Biology Kingston (2002), G4,6,8 PEM (2002), Earth Science PEM (2002), Biology Pommerich (2004), G11,12 Russell (1999), G8 Way, Davis, & Fitzpatrick (2006), G11
Social Studies			Kingston (2002), G4,6,8 Way, Davis, & Fitzpatrick (2006), G8,11

Student preferences

Some studies asked students who took computerized tests whether they would prefer to take future tests on computer or on paper. In all such studies located for this review, the majority of students indicated their preference to test on computer (Bridgeman, Lennon, & Jackenthal, 2001; Higgins, Russell, & Hoffman, 2005; Glassnapp, Poggio, Poggio, & Yang, 2005; Ito & Sykes, 2004; Johnson & Green, 2004; O'Malley et al., 2005; Richardson et al., 2002; Sim & Horton, 2004).

Comparability for different student subgroups

While equity is a critical concern, for a variety of logistical and cost reasons, most studies do not focus on comparability for different subgroups of students. Some findings from studies that have looked at these issues follow.

Computer experience. While some early studies suggested that students who had less experience with computers would score lower on computer-administered tests, recent studies find no evidence of such a disadvantage (Bennett, 2002; Higgins, Russell, & Hoffman, 2005).

Race/ethnicity. Ewing, Wiley, & Gillie (2003) found computer-based and paper-and-pencil math tests had the same factor structure for African-American, Asian, and Hispanic students.

That is, for each ethnic group the same pattern of sub-scores emerged for a given total score. But they found differences for English composition. That is, the pattern of sub-scores (for a given total score) for African-American, Asian, and Hispanic students tended to vary.

Nichols and Kirkpatrick (2005) found no differences in administration mode comparability among various demographic subgroups.

Gender. Sim & Horton (2005) did not find any comparability differences based on gender. McCann (2006), in an analysis of Australian students, also failed to find any gender effect.

Impact of monitor quality

Several authors have indicated that the quality of the presentation of text online (as related to monitor size and resolution) can negatively impact comparability (Dyson & Kipling, 1997; Schenkman, Fukada, & Perrson, 1999).

Impact of network quality

Anecdotal evidence from several researchers suggests that slow response times, whether due to internal school network constraints or varying speed of internet connections, can frustrate students (and teachers) during online testing and can negatively impact student performance.

Impact of test speededness

In a summary of 28 studies of 159 tests, Mead and Drasgow (1993) noted that comparability was most greatly impacted in speeded tests (tests which, usually purposely, do not provide sufficient time for all examinees to finish). The studies they examined were of adult populations. One recent study found similar results for students in Grades 4–12 taking a cognitive abilities test (Ito & Sykes, 2004). It should be noted that the FCAT retest is not a speeded test.

Summary

The preponderance of studies of the comparability of K–12 computer- and paper-administered multiple-choice tests has shown differences that are either statistically not significant or of no practical significance. However, other studies have shown advantage to either paper or computer administration. Each year, more studies are being conducted, and our understanding of potential differences in the testing modes is increasing. In the meantime, we hope this information will help you decide whether your students should take paper or computerized versions of the Sunshine State Standards (SSS) Reading and the SSS Mathematics Grade 10 FCAT.

References and bibliography

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.

Bennett, R.E. (2002). Using electronic assessment to measure student performance. 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. Educational Testing Service: Princeton, NJ

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). Educational Testing Service, Princeton, NJ

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 annual meeting of American Education Research Association, San Diego, CA

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

Dyson, M. C. & Kipping, G. J. (1997). Legibility of screen formats: are three columns better than one? Computers & Graphics, 21(6), 703–712.

Ewing, M., Wiley, A., & Gillie, J.M. (2003). Moving from paper-and-pencil administration to computer-based testing: an investigation of construct equivalence and subgroup differences. Paper presented at annual meeting of National Council on Measurement in Education, Chicago, IL

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

Glassnapp, D.R., Poggio, J., Poggio, A., & Yang, X. (2005). Student attitudes and perceptions regarding computerized testing and the relationship to performance in large-scale assessment programs. Paper presented at annual meeting of National Council on Measurement in Education, Montreal, Canada

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

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

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

Johnson, M. & Green, S. (2004). On-line assessment: the impact of mode on students' strategies, perceptions, and behaviours. Paper presented at annual meeting of British Educational Research Association, Manchester, Great Britain

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

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

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

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

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

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

O'Malley, K. J., Kirkpatrick, R., Sherwood, W., Burdick, H. J., Hsieh, M.C., Sanford, E.E. (2005). Comparability of a paper-based and computer-based reading test in early elementary grades. Paper presented at AERA Division D Graduate Student Seminar, Montreal, Canada.

Paek, P. (2005). Recent trends in comparability studies (PEM Research Report 05-05). Available from http://www.pearsonedmeasurement.com/downloads/research/RR_05_05.pdf

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

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

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

Poggio, J., Glasnapp, D.R., Yang, X., & Poggio, A.J. (2005). Comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. 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. Journal of Technology, Learning, and Assessment, 2(6), 1–44.

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

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

Richardson, M., Baird, J., Ridgway, J., Ripley, M., Shorrocks-Taylor, D., & Swan, M. (2002). Challenging Minds? Students' perceptions of computer-based World Class Tests of problem-solving. Computers and Human Behavior, 18(6), 633–49

Russell, M. (1999). Testing on computers: a follow-up study comparing performance on computer and on paper. Education Policy Analysis Archives, 7(20). Available online 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). Available online 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). Available online 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. Available online from http://www.tcrecord.org

Sandene, Bennett, Braswell, & Oranje (2005). Online Assessment in Mathematics and Writing: Reports From the NAEP Technology-Based Assessment Project, Research and Development Series. Available online from http://nces.ed.gov/nationsreportcard/pubs/studies/2005457.asp

Schenkman B., Fukada T, & Persson B. (1999). Glare from monitors measured with subjective scales and eye movements, Displays, vol. 20: 11–21.

Sim, G. & Horton, M. (2005). Performance and attitude of children in computer based versus paper based testing. Available at http://www.uclan.ac.uk/facs/destech/compute/staff/read/Publish/ChiCi/references/performance_and_attitude.pdf

Wang, S. (2004). Online or paper: does delivery affect results? Administration mode comparability study for Stanford diagnostic Reading and Mathematics tests. San Antonio, TX

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.

Way, D., Davis, L., & Fitzpatrick, S. (2006). Score comparability of online and paper administrations of the Texas Assessment of Knowledge and skills. Paper presented at annual meeting of National Council on Measurement in Education, San Francisco, CA