



TECHNOLOGY AND ASSESSMENT STUDY COLLABORATIVE

Bridging the Gap Between Testing and Technology in Schools

Michael Russell & Walt Haney
Technology and Assessment Study Collaborative
Boston College
332 Champion Hall
Chestnut Hill, MA 02467

www.intasc.org



Bridging the Gap Between Testing and Technology in Schools

Michael Russell and Walt Haney
Technology and Assessment Study Collaborative
CSTEED, Boston College
Rereleased by inTASC July 2002
Originally published April 2000 in Education Policy Analysis Archives, 8(19),
<http://epaa.asu.edu/epaa/v8n19.html>

The need for improving education in the U.S. has received unprecedented attention recently in the media and in national and state elections. Prescriptions for improving schools have been many, but two of the most common are what might be called the technology and testing remedies.

The technology nostrum holds that the infusion of modern technology into schools will bolster teaching and learning and will prepare students for an increasingly technological workplace. The second prescription, which is often called high stakes testing, holds that standards-based accountability for students, teachers and schools will provide tangible incentives for improvements in teaching and learning. What is little recognized, however, is that these two strategies are working against each other in a sort of educational time warp. Recent research shows that written tests taken on paper severely underestimate the performance of students accustomed to working on computer (Russell, 1999; Russell & Haney, 1997). The situation is analogous to testing the accounting skills of modern accountants, but restricting them to the use of an abacus for calculations.

The widening gap between the increased use of technology in schools and the absence of computers in state-level testing programs raises important implications for policies related to the use of both technology and testing in schools. In this article, we summarize recent developments in the use of technology in schools and in state level testing programs. We then describe two studies indicating that written tests administered on paper underestimate the achievement of students accustomed to working on computers. We conclude by discussing four approaches to bridging the gap between technology and testing in US schools.

The Computer Revolution Goes to School

Although the personal-computer revolution began only twenty years ago and widespread use of the world wide web (WWW) is even more recent, computer technology has already had a dramatic impact on society and schooling. Between 1984 and 1993, the percentage of people using computers in the workplace nearly doubled from 24.6 percent to 45.8 percent. Similarly, the percentage of people owning one or more computers in their home increased rapidly from 8.2 percent in 1984 to 22.8 percent in 1993 to 33.6 percent in 1997 (Newburger, 1997). Although schools have been slower to acquire these technologies, computer use in schools has recently increased rapidly (Zandvliet & Farragher, 1997). While schools had one computer for every 125 students in 1983, they had one for every 9 students in 1995, and 1 for every 6 students in 1998 (Market Data Retrieval, 1999).

Not only are more computers in classrooms, but schools are also increasing students' use of computers and access to the Internet. A recent national survey of teachers showed that in 1998, 50 percent of K–12 teachers had students use word processors, 36 percent had them use CD ROMS, and 29 percent had them use the WWW (Becker, 1999). Although it is unclear how computers are affecting student achievement in schools (see, for example, Fabos & Young, 1999, questioning the efficacy of Internet based telecommunications exchange programs in schools), there is little doubt that the computer revolution has gone to school. As a result, more and more students are writing and performing school assignments on computers.

Performance Testing in Schools

Meanwhile, many states are increasingly seeking to hold students, teachers and schools accountable for student learning as measured by state-sponsored tests. According to annual surveys by the Council for Chief State School Officers (CCSSO, 1998), 48 states use statewide tests to assess student performance in different subject areas. Many of these tests are tied to challenging standards for what students should know and be able to do. Scores on these tests are being used to determine whether to: (1) promote students to higher grades; (2) grant high school diplomas; and (3) identify and sanction or reward low- and high-performing schools (Sacks, 1999). Currently, 32 states control, or plan to control, graduation and/or grade promotion via student performance on state-level tests. Because of the limitations of multiple-choice tests, many statewide tests include sections in which students must write extended answers or written explanations of their work. As the recent CCSSO report commented, "Possibly the greatest changes in the nature of state student assessment programs have taken place in the 1990s as more states have incorporated open-ended and performance exercises into their tests, and moved away from reliance on only multiple-choice tests" (CCSSO, 1998, p. 17). In 1996–97, an estimated ten to twelve million students nationwide participated in a state-sponsored testing program that required them to write responses long hand (given a total national K–12 enrollment of about 50 million and open-ended assessments in almost all the states in 3 out of 12 grade levels). In Ohio, for example, students must pass the written portion of the Ohio Proficiency Test in order to graduate from high school (Fisher & Elliott, 2000).

Although many observers have criticized state-sponsored high-stakes tests on a variety of grounds (e.g., Heubert & Hauser, 1999; Sacks, 1999), here we direct attention to a widely unrecognized but critical limitation of depending on these tests to drive educational reform—paper-and-pencil written tests yield misleading information on the capabilities of students accustomed to using computers.

Testing Via Computer

Research on testing via computer goes back several decades and suggests that for multiple-choice tests, administration via computer yields about the same results, at least on average, as administering tests via paper-and-pencil (Bunderson, Inouye, & Olsen, 1989, Mead & Drasgow, 1993). However, more recent research shows that for young people who have gone to school with computers, open-ended (that is, not multiple choice) questions administered via paper-and-pencil yield severe underestimates of students' skills as compared with the same questions administered via computer (Russell, 1999; Russell & Haney, 1997). In both studies, the effect sizes for students accustomed to working computer ranged from .57 to 1.25. Effect sizes of this magnitude are unusually large and of sufficient size to be of not just statistical, but also practical significance (Cohen, 1988; Wolf, 1986). Effect sizes of this magnitude, for example, imply that the score for the average student in the experimental group tested on computer exceeds that of 72 to 89 percent of the students in the control group tested via paper and pencil.

Our research on this topic began with a puzzle. While evaluating the progress of student learning in the Accelerated Learning Laboratory (ALL), a high-tech school in Worcester, MA, teachers were surprised by the results from the second year of assessments. Although students wrote more often after computers were widely used in the school, student scores on writing tests declined in the second year of the new program. To help solve the puzzle, the school asked us to assist in comparing paper and computer administration of the tests.

In 1995, a randomized experiment was conducted, with one group of sixty-eight students taking math, science and language arts tests, including both multiple-choice and open-ended items, on paper, and another group of forty-six students taking the same tests on computer (but without access to word processing tools, such as spell-checking or grammar-checking). Before scoring, answers written by hand were transcribed so that raters could not distinguish them from those done on computer. There were two major findings. First, the multiple-choice test results did not differ much by mode of administration. Second, the results for the open-ended tests differed significantly by mode of administration. For the ALL School students who were accustomed to writing on the computer, responses written on computer were much better than those written by hand. This finding occurred across all three subjects tested and on both short answer and extended answer items. The effects were so large that when students wrote on paper, only 30 percent performed at a "passing" level; when they wrote on computer, 67 percent "passed" (Russell & Haney, 1997).

Two years later, a more sophisticated study was conducted, this time using open-ended items from the new Massachusetts state test (the Massachusetts Comprehensive Assessment System or MCAS) and the National Assessment of Educational Progress

(NAEP) in the areas of language arts, science and math. Again, eighth grade students from two middle schools in Worcester, MA, were randomly assigned to groups. Within each subject area, each group was given the same test items, with one group answering on paper and the other on computer. In addition, data were collected on students' keyboarding speed and prior computer use. As in the first study, all answers written by hand were transcribed to computer text before scoring.

In the second study, which included about two hundred students, large differences between computer and paper-and-pencil administration were again evident on the language arts tests. For students who could keyboard moderately well (20 words per minute or more), performance on computer was much better than on paper. For these students, the difference between performance on computer and on paper was roughly a half standard deviation. According to test norms, this difference is larger than the amount students' scores typically change between grade 7 and grade 8 on standardized tests (Haney, Madaus, & Lyons, 1993, p. 234). For the MCAS, this difference in performance could easily raise students' scores from the "failing" to the "passing" level (Russell, 1999).

Recalling that nearly ten million students took some type of state-sponsored written test last year and that nearly half of the students nationwide use word processors in school, these results suggest that state paper-and-pencil tests may be underestimating the abilities of millions of students annually.

In the second study, however, findings were not consistent across all levels of keyboarding proficiency. As keyboarding speed decreased, the benefit of computer administration became smaller. And at very low levels of keyboarding speed, taking the test on computer diminished students' performance (effect size of about 0.40 standard deviations). Similarly, taking the math test on computer had a negative effect on students' scores. This effect, however, became less pronounced as keyboarding speed increased.

Bridging the Gap

These studies highlight the importance of the gap between the technology and testing strategies for school improvement. Increasingly, schools are using computers to improve student learning. To measure increases in student learning, states are depending upon tests administered on paper. The open-ended questions on these tests, however, underestimate the achievement of students who regularly use computers. As a result, this mis-match between the mode of learning and the mode of assessment may be underestimating improvements in achievement. This problem is likely to increase as more students become accustomed to writing on computers. There are at least four possible ways to bridge this gap.

First, schools could decrease the amount of time students spend working on computers so that they do not become accustomed to writing on computers. Some schools have already adopted this practice. After reviewing the first study described above and following the introduction of the new paper-and-pencil MCAS test in Massachusetts, the ALL school required students to write more on paper and less on computer (Russell, 1999). In another Massachusetts school system, the principal feared that students who write regularly on computer lose penmanship skills, which might lead

to lower scores on the new state test. This school increased penmanship instruction across all grades while also decreasing students' time on computers (Holmes, 1999). Such strategies, in effect reducing computer use in schools to better prepare students for low-tech tests, may be pragmatic given the high stakes attached to many state tests. But they are also short-sighted in light of students' entry after graduation into an increasingly high-tech world and workplace.

A second way to bridge the test-technology gap would be to eliminate paper-and-pencil testing and have students perform open-ended tests on computer. This might seem a sensible solution, but it will not be feasible until all schools obtain an adequate technology infrastructure. Moreover, as shown by problems in recent moves to administer some large-scale tests for adults on computers, computerized testing is not the panacea some had hoped. Among other problems, it adds considerably to the cost of testing and creates new test security concerns. But more importantly, as the second study summarized above indicates, administering open-ended tests only on computer would penalize students with poor keyboarding skills.

A third approach would be to offer students the option of performing open-ended tests on paper or on computer. On the surface, this seems like a sensible solution. However, it would add considerable complexity and cost to test administration and scoring procedures. Although there has not been a large amount of research on the extent to which computer printing versus hand-writing affects ratings of written work, Powers et. al. (1994) report that significant effects can occur. Surprisingly, Powers et. al. found that computer printed responses produced by adults tended to receive lower scores than the same responses produced by hand. To control for such effects, in offering tests on paper and computer, handwritten responses would need to be converted to computer text. Surely it will be some years before text recognition software is sophisticated enough to convert handwritten responses into computer text. Thus, for the foreseeable future, the cost of transcription would be prohibitive.

But beyond the need to convert responses to the same medium for scoring, the second study summarized above provides evidence that, when given the choice of using computer or paper to write their tests, many students make poor decisions as to which medium they should use. This was evidenced in two ways. First, the correlations between both students' preference for taking tests on computer or on paper and keyboarding speed and between preference and prior computer experience were near zero (less than .18). Second, preference was not found to be a significant factor in predicting students performance. Together, the added complexity of scoring open-ended responses produced in both mediums and students' apparent inaccuracy in selecting the medium that optimizes their performance suggest that simply giving students the option of performing open-ended tests on computer or on paper would do little to reduce the gap between testing and technology.

A fourth approach, and perhaps the most reasonable solution in the short term, is to recognize the limitations of current testing programs. Without question, both computer technology and performance testing can help improve the quality of education. However, until students' can take tests in the same medium in which they generally work and learn, we must recognize that the scores from high-stakes state tests do not accurately reflect some students' capabilities. Reliance on paper and pencil written

test scores to measure or judge student and/or school performance will mischaracterize the achievement of students' accustomed to working on computers. Thus, the gap between the use of technology in schools and testing programs serves as yet another reminder of the dangers of judging students and schools based solely on written test scores.



Acknowledgments

We would like to acknowledge the help of Jeff Nellhaus and Kit Viator of the Massachusetts Department of Education which allowed inclusion of MCAS items in the second study summarized in this article. Also, we wish to note that this article is an expansion of an opinion essay appearing originally in the *Christian Science Monitor* in July 1999 (Haney & Russell, 1999). Additionally, we thank the National Board on Educational Testing and Public Policy (NBETPP) for its support and suggestions on earlier versions of this article. We also thank two anonymous EPAA reviewers for their suggestions for improving this article. Finally we thank Carol Shilinsky and the staff of the ALL School, and James Caradonio, the Superintendent of the Worcester, MA, Public Schools, for their generous support of the research recounted here.

References

- Becker, H. J. (1999). *Internet Use by Teachers: Conditions of Professional Use and Teacher-Directed Student Use*. Irvine, CA: Center for Research on Information Technology and Organizations.
- Bunderson, C. V., Inouye, D. K. & Olsen, J. B. (1989). The four generations of computerized educational measurement. In Linn, R. L., *Educational Measurement* (3rd ed.), Washington, D.C.: American Council on Education, pp. 367–407.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.) Hillsdale, NJ: Erlbaum.
- Council of Chief State School Officers (1998). *Key State Education Policies on K-12 Education: Standards, Graduation, Assessment, Teacher Licensure, Time and Attendance*. Washington, DC.
- Fabos, B. & Young, M. (1999). Telecommunications in the classroom: Rhetoric versus reality. *Review of Educational Research*. 69:3 217–259.
- Fisher, M. & Elliott, S. (2000) Proficiency: The test questioned. Dayton Daily News (March 12). <http://www.activedayton.com/news/2000/03/12/0313main>.
- Glennan, T. K., & Melmed, A. (1996). *Fostering the use of educational technology: Elements of a national strategy*. Santa Monica, CA: RAND.
- Haney, W., Madaus, G., & Lyons, R. (1993). *The Fractured Marketplace for Standardized Testing*. Boston, MA: Kluwer Academic Publishers.
- Haney, W. & Russel, M. (1999). Low-tech tests shortchange high-tech students. *Christian Science Monitor*. July 1, 1999.
- Holmes, R. (1999). A gender bias in the MCAS? MetroWest Town Online. <http://www.townonline.com/metrowest/archive/022499/>.
- Heubert, J. & Hauser, R. (1999). *High stakes: Testing for tracking, promotion and graduation*. (Report of the Committee on Appropriate Test Use). Washington, D.C.: National Academy Press. (Available on-line at <http://www.nap.edu>).
- Market Data Retrieval. (1999). *Technology in Education 1999*. (A report issued by Market Data Retrieval). Shelton, CN: Market Data Retrieval.

- Mead, A. D. & Drasgow, (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114:3, 449–58.
- Newburger, E. C. (1997). *Computer Use In the United States*. Washington, DC: U.S. Census Bureau.
- Powers, D., Fowles, M, Farnum, M, & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement*, 31(3), 220–233.
- 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(1), <http://epaa.asu.edu/epaa/v5n3.html>.
- Russell, M. (1999). Testing Writing on Computers: A Follow-up Study Comparing Performance on Computer and on Paper. *Educational Policy Analysis Archives*, 7(20), <http://epaa.asu.edu/epaa/v7n20/>.
- Sacks, P. (1999). *Standardized minds*. Reading MA: Perseus Books.
- Snyder, T. D. & Hoffman, C. (1990). *Digest of Education Statistics*. Washington, DC: U. S. Department of Education.
- Snyder, T. D. & Hoffman, C. (1994). *Digest of Education Statistics*. Washington, DC: U. S. Department of Education.
- 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.
- Zandvliet, D. & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education*, 29(4), 423–438.

