

Lynch School of Education

inTASC Publications

Boston College

Year 2003

Computer-Based Testing and Validity: A Look Back and Into the Future

Michael Russell
Boston College,

Kathleen O'Connor
Boston College,



TECHNOLOGY AND ASSESSMENT STUDY COLLABORATIVE

Computer-Based Testing and Validity: A Look Back and Into the Future

Michael Russell, Amie Goldberg, & Kathleen O'Connor
Technology and Assessment Study Collaborative
Boston College
332 Champion Hall
Chestnut Hill, MA 02467

www.intasc.org

Computer-Based Testing and Validity: A Look Back and Into the Future

Michael Russell, Amie Goldberg, & Kathleen O'Connor
Technology and Assessment Study Collaborative
Boston College
Released August 2003

Michael K. Russell, Project Director/Boston College

Copyright © 2003 Technology and Assessment Study Collaborative, Boston College

Supported under the Field Initiated Study Grant Program, PR/Award Number R305T010065, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the Office of Educational Research and Improvement, or the U.S. Department of Education.



Computer-Based Testing and Validity: A Look Back and Into the Future

Michael Russell, Amie Goldberg, & Kathleen O'Connor
Technology and Assessment Study Collaborative
Boston College

Released July 2003

Over the past decade, access to and use of computers in homes and schools have increased sharply. Both in and out of the classroom, students' educational use of computers has also increased, particularly for writing and research (Becker, 1999; Russell, O'Brien, Bebell, & O'Dwyer, 2002). Over this same time period, reliance on large-scale tests to compare the quality of education provided by nations has increased. For example, every four years, the Trends in International Mathematics and Science Study compares the performance of approximately 40 nations in mathematics and science. Similarly, the Progress in International Reading Literacy Study compares literacy achievement across 35 nations. In the United States, the use of large-scale tests to make decisions about the quality of education provided by individual schools and the achievement of individual students has also exploded. For example, the number of U.S. states that have developed tests linked to standards has increased steadily from zero in 1983 to 37 in 2001 (Meyer, Orlofsky, Skinner & Spicer, 2002, p. 68). Similarly, the number of states that seek to hold schools and students accountable by requiring students to pass high school graduation tests has risen steadily from four in 1983 to an expected 27 by 2008 (Amrein & Berliner, 2002).

Despite steady growth in the access to and use of computers for classroom learning and the rapid increase in the use of tests to make decisions about national educational programs, schools, and students, the use of computers for elementary, middle, and secondary school testing programs was absent until the turn of the century. Since then, the number of U.S. states exploring or using computer-based tests has increased rapidly. As of this writing, at least 12 state testing programs have begun exploring the use of computers. Similarly, nations like Singapore and Norway are beginning to consider ways in which computers might be used to enhance student assessment. Although it is too early to gauge the success of these programs, it is likely that other states and

nations will also begin transitioning their tests to a computer format over the next few years.

As elementary and secondary school-level testing programs begin administering tests on computers, several issues related to the validity of using test scores provided by computer-based tests to make decisions about student and school performance must be examined. In this paper, we describe how efforts to examine the validity of computer-based test (CBT) scores have evolved over the past thirty years. We then discuss how variations in students' use of computers for writing complicate efforts to obtain valid measures of student writing skills, regardless of test mode. In this discussion we explore aspects of validity, namely construct and consequences, which may be threatened by paper or computer-based tests. We conclude by discussing the conflict caused by the traditional desire to deliver tests in a standardized format. During this discussion, we argue that while standardization is necessary for norm-referenced tests, it may be less applicable for the standards-based criterion-referenced tests used by many state testing programs.

Computerized Testing: Early Era, 1969 Through 1985

In the early seventies, the U.S. military and clinical psychologists pioneered the development of computer-based tests. Initially, psychologists saw computerized assessments as a method for controlling test variability, eliminating examiner bias, and increasing efficiency. With computerized testing, psychologists could optimize the use of trained personnel by freeing them from the routine functions of test administration and scoring.

Between 1970 and 1985, comparative studies of conventional and computer-based test formats were conducted for a variety of test instruments including personality assessments, intelligence and cognitive ability scales, and vocational interest inventories. These comparisons were performed to examine the equivalence of administering the same set of items on paper or on a computer. Thus, the question addressed by this line of research focused on the interchangeability of scores obtained from a paper or computer-based test. In these cross-modal validity studies, test equivalence was established when group performance did not differ significantly between modes.

In general, evidence for cross-modal validity was found for self-report instruments, such as the Minnesota Multiphasic Personality Inventory (Bisken & Kolotkin, 1977; Bresolin, 1984; Evan & Miller, 1969; Koson, Kitchen, Kochen, & Stodolosky, 1970; Lushene, O'Neil, & Dunn, 1974; White, Clements, & Fowler, 1985) and cognitive batteries/intelligence scales, such as the Wechsler Adult Intelligence Scale (Elwood, 1969; Elwood & Griffin, 1972) for which cross-modal correlations of .90 or higher were found for all subtests.

Studies that focused on personality inventories provided similar results. As an example, Katz and Dalby (1981a) found no significant differences in performance between forms of the Eysenck Personality Inventory with mental

health patients. Hitti, Riffer, and Stuckless (1971), and later, Watts, Baddeley, and Williams (1982) reported only slight differences between computer-based test scores and paper-and-pencil scores for the Raven Progressive Matrices, while Scissons (1976) found “notable”, but not significant differences in scores between forms of the California Psychological Inventory.

In the field of education, few comparative studies on test mode were carried out prior to 1986. However, an early 1985 study by Lee and Hopkins found the mean paper-and-pencil test score to be significantly higher than the mean computer-based test score in arithmetic reasoning. Results of this study highlight scratchwork space as a salient factor in arithmetic test performance. In addition, Lee and Hopkins concluded that the inability to review and revise work affected performance, and argued that only “software that allows the conveniences of paper-and-pencil tests, e.g., the ability to change answers and the ability to review past items, be used in future applications” (p. 9). Collectively, early research on cross-modal validity of arithmetic reasoning tests provided mixed results: computers were found to enhance (Johnson & Mihal, 1973), as well as impede (Lee, Moreno, & Sympson, 1986) test performance.

During the early era of research on the validity of computer-based tests, one study focused specifically on a special population of examinees. In a study conducted by Katz and Dalby (1981b), results were reported separately for children labeled as “gifted” or as “behavioral problems”. For both groups, Katz and Dalby (1981b) found evidence for cross-modal validity (i.e., no significant differences in test scores between modes) on the Fundamental Interpersonal Relations Orientation (FIRO-BC), an instrument designed to measure children’s characteristic behavior toward other children.

While findings from many of these early investigations were encouraging for computer-based test advocates, many of the studies did not employ rigorous methodologies that would indicate true test equivalency. In addition, these studies often used convenience samples consisting of undergraduate students from a particular university. As a result, it was not possible to generalize from the study sample to the population of test takers for which the instrument was designed.

Additionally, prior to the mid-1980s, computer use was not yet widespread. For this reason, only a small portion of the population was adept with computers. For many test-takers, using a computer for any purpose, particularly to take a test, was a novel experience. As a result, it was difficult to disentangle the positive and negative effects of using a computer for testing. In some cases, differences in performance between modes of administration could be attributable to anxiety and/or computer illiteracy (Hedl, O’Neil, & Hansen, 1973; Johnson & White, 1980; Lee, Moreno, & Sympson, 1986). At times, technical difficulties resulting from the under-developed technology and cumbersome interfaces interfered with an examinee’s performance. In other cases, the use of computers by one group may have elevated effort on the test.

Despite these limitations, some important factors regarding validity with computerized testing began to emerge. Specifically, administration factors,

such as the transfer of problems from the screen to scratchwork space, lack of scratchwork space, and inability to review and/or skip individual test items, were found to significantly affect test performance.

Computer-Based Testing Guidelines: The American Psychological Association's (APA) Guidelines for Computer-Based Tests and Interpretations, 1986

With the development of computerized tests burgeoning, the American Psychological Association (APA) released a preliminary set of guidelines, titled *Guidelines for Computer-Based Tests and Interpretations* (APA, 1986). This document contains specific recommendations for computerized test administrations and score interpretations. With respect to computerized test design, for example, the guidelines state the “computerized administration normally should provide test takers with at least the same degree of feedback and editorial control regarding their responses that they would experience in traditional testing formats” (APA, 1986, p. 12). In other words, normally, test takers should be able to review their responses to previous items as well as skip ahead to future items, and make any changes they wish along the way.

In addition to guidelines related to human factors, there are specific guidelines that address relevant psychometric issues. Guidelines 16 and 19 state that the equivalence of scores from computerized and conventional testing should be reported to establish the relative reliability of computerized assessment. The guidelines further elaborate that computer and conventional administration formats are generally considered equivalent if they satisfy three criteria: the same mean scores, standard deviations, and rankings of individual examinees. Additionally, APA's guidelines state that evidence of equivalence between two forms of a test must be provided by the publisher and that the users of computer-based tests be made aware of any inequalities resulting from the modality of administration.

Computer-Based Assessment Validity Studies: 1986 (Post APA Guidelines) to Present

In contrast to validity studies conducted during the first 15 years of computer-based testing, in the later part of the 1980's, researchers began focusing on features idiosyncratic to the computer format in order to identify key factors that affect examinee performance. Among the factors studied were a) the ability to review and revise responses, b) the presentation of graphics and text on computer screens, and c) prior experience working with computers.

Reviewing and Revising Responses

Building on Lee and Hopkin's (1985) earlier work, several studies provided confirmatory evidence that the inability to review and revise responses had

a significant negative effect on examinee performance (Wise & Plake, 1989; Vispoel, Wang, de la Torre, Bleiler, & Dings, 1992). As an example, Vispoel et al. (1992) found the ability to review work modestly improved test performance, slightly decreased measurement precision, moderately increased total testing time, and was strongly favored by examinees of a college level vocabulary test. These findings led Vispoel et al. to conclude that computerized tests with and without item review do not necessarily yield equivalent results, and that such tests may have to be equated to ensure fair use of test scores.

It is interesting to note that the effect of item review on examinee performance is part of a larger issue concerning the amount of control that computer-based testing examinees should be provided. Wise and Plake (1989) noted there are three basic features available to examinees during paper-and-pencil tests that should be considered by computer-based testing developers: allowing examinees to skip items and answer them later in the test, allowing examinees to review items already answered, and allowing examinees to change answers to items.

Although the effects of denying the computer-based testing examinee such editorial controls have not been fully examined, over the past fifty years, a substantial body of research has examined the effect of answer changing on test performance within the context of paper-and-pencil tests. Summarizing this body of research, Mueller and Wasser (1977) report gain to loss ratios for multiple-choice items range from 2.3:1 to 5.3:1. These findings indicate that for every one answer change that results in an incorrect response, there are over two, and in some cases over five, answer changes that result in correct responses. Mueller and Wasser concluded that students throughout the total test score distribution gain more than they lose by changing answers, although higher scoring students tend to gain more than do lower scoring students. And, when total test score is controlled for, there appears to be no difference in score gain between genders (Mueller & Wasser, p. 12). Such findings suggest that item review is an important test-taking strategy that has a positive effect on examinee performance.

Item Layout and Presentation of Graphics

A second line of research examined the effect of item layout on examinee performance. As an example, Mazzeo and Harvey (1988) found that tests that require multiscreen, graphical, or complex displays result in modal effects. In addition, graphical display issues such as the size of the computer screen, font size, and resolution of graphics were found to affect examinee performance. In such cases, some researchers argued that these and other computer-linked factors may “change the nature of a task so dramatically that one could not say the computerized and conventional paper-and-pencil version of a test are measuring the same construct” (McKee & Levinson, 1990, p. 328).

In response to this growing body of research, Mazzeo and Harvey (1988) urged test publishers to conduct separate equating and norming studies when

introducing computerized versions of standardized tests. In a later study, Mazzeo, Druesne, Raffeld, Checketts, and Muhlstein (1991) reaffirmed the need to determine empirically the equivalence of computer and paper versions of an examination. Additionally, Kolen and Brennan (1995) argue that mode effects of paper-and-pencil and computer-based tests are complex and that the extent to which effects are present are likely dependent on the particular testing program. This complexity also suggests that separate analyses for modal sensitivity are necessary for any test offered in both formats. Finally, Mazzeo et al. (1991) recommended examining relative effects of mode of administration among subpopulation groups, which, over a decade later, is still not a factor commonly studied.

Comfort with Computers

A third line of research investigated the effect of comfort and familiarity with computers on examinee performance. Predictably, much of this research indicates that familiarity with computers does play a significant role in test performance (Ward, Hooper, & Hannafin, 1989; Llabre, Clements, Fitzhugh, Lancelotta, Mazzagatti, & Quinones, 1987). It is important to note, however, that much of this research was conducted prior to the widespread penetration of computers into homes, schools, and the workplace and before graphical user interfaces were used widely. Therefore, in much of this research, examinees' comfort levels with computers were low. More recent research on comfort, familiarity, and examinee performance has focused largely on assessing student writing ability (Russell & Haney, 1997; Russell, 1999; Russell & Plati, 2001; 2002). As described in greater detail below, this body of research suggests that not only are some students more comfortable and accustomed to writing on computers but computer-based testing may provide a better mode than traditional paper-and-pencil testing for assessing their writing ability.

In the following section we will focus on this complex interplay of increased computer use, student comfort, and valid assessment modes in the context of measuring student performance.

Student Computer Use, Writing, and State Testing Programs

As summarized above, a substantial body of research has examined the equivalence of scores provided by tests administered on paper and on computer. In most of the research examining the comfort level examinees have with computers, the focus is on how familiarity with computers affects the performance of examinees when they take a test on computer. In some cases, research has found that examinees who are less familiar with computers perform worse when they take a test on a computer. In such cases, test developers may be inclined to administer the test on paper so as not to disadvantage examinees who are unfamiliar with computers.

Increasingly, however, computers are used in schools to develop student

skills and knowledge. As one example, computers are used by a large percentage of students to develop writing skills (Becker, 1999; Russell, O'Brien, Bebell, & O'Dwyer, 2003). Despite regular use of computers by many students to produce writing, state testing programs currently require students to produce responses to open-ended and essay questions using paper-and-pencil. However, as mentioned above, there is increasing evidence that tests that require students to produce written responses on paper underestimate the performance of students who are accustomed to writing with computers (Russell & Haney, 1997; Russell, 1999; Russell & Plati, 2001, 2002).

In a series of randomized experiments, this mode of administration effect has ranged from an effect size of about .4 to just over 1.0. In practical terms, the first study indicated that when students accustomed to writing on computer were forced to use paper-and-pencil, only 30% performed at a "passing" level; when they wrote on computer, 67% "passed" (Russell & Haney, 1997). In a second study, for students who could keyboard approximately 20 words a minute, the difference in performance on paper versus on computer was larger than the amount students' scores typically change between seventh and eighth grade on standardized tests. However, for students who were not accustomed to writing on computer and could only keyboard at relatively low levels, taking the tests on computer diminished performance (Russell, 1999). Finally, a third study that focused on the Massachusetts Comprehensive Assessment System's (MCAS) Language Arts Tests demonstrated that removing the mode of administration effect for writing items would have a dramatic impact on the study district's results. As Figure 1 indicates, based on 1999 MCAS results, 19% of the fourth graders classified as "Needs Improvement" would move up to the "Proficient" performance level. An additional 5% of students who were classified as "Proficient" would be deemed "Advanced" (Russell & Plati, 2001, 2002).

Figure 1

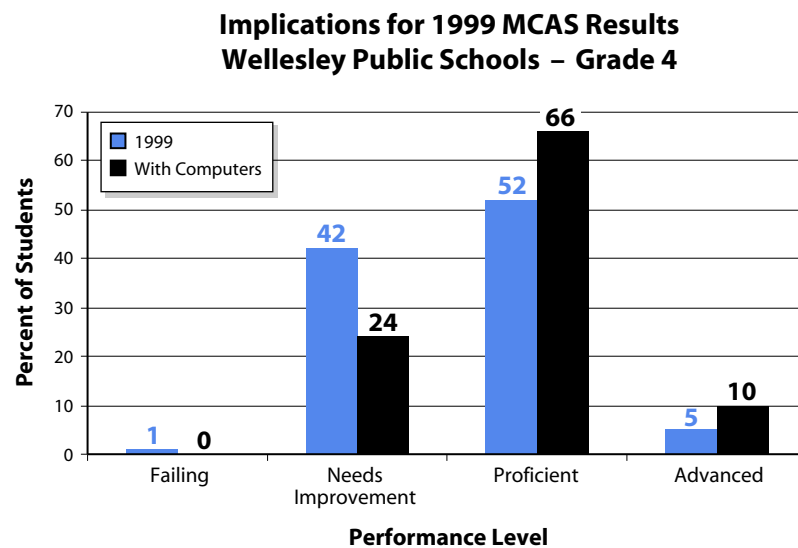


Figure 1. Mode of administration effect on grade 4 MCAS results (from Russell & Plati, 2001).

In essence, this body of research provides evidence that cross-modal validity does not exist for these writing tests. Depending on the mode of administration, the performance of some students is mis-measured. Specifically, the performance of students who are not accustomed to writing with computers is underestimated by computer-based writing tests. Conversely, students who are accustomed to writing with computers are mis-measured by paper-based tests. In essence, this problem results from a mixing of two constructs: a) written communication skills and b) text production skills. For students accustomed to writing on paper, inability to use the keyboard to efficiently record and edit ideas interferes with their ability to communicate their thinking in writing. Conversely, for students accustomed to writing on computers, recording ideas on paper by hand interferes with their ability to fluidly record and edit their ideas.

Beyond mismeasuring the performance of students, recent evidence indicates that the mode of administration effect may also have negative consequences on instructional uses of computers. Russell and Haney (2000) provided two lines of evidence that teachers in two schools had already begun to reduce instructional uses of computers so that students do not become accustomed to writing on computers. In one case, following the introduction of the new paper-and-pencil test in Massachusetts, the Accelerated Learning Laboratory (a K–8 school in Worcester that was infused with computer-based technology) required students to write more on paper and less on computer. Fearing that students who write regularly with a computer might lose penmanship skills, a principal in another school increased the amount of time teachers spent teaching penmanship and decreased the amount of time students wrote using computers.

More recently, a national survey of teachers conducted by the National Board on Educational Testing and Public Policy (Pedulla, Abrams, Madaus, Russell, Ramos, & Miao, 2003) provides insight into the ways in which teachers believe they are changing their instructional practices in response to state-level testing programs. Among the several Likert response questions administered to a representative sample of over 4,000 teachers nationwide during the winter of 2001, two focused specifically on instructional uses of technology. The items were as follows:

- Teachers in my school do not use computers when teaching writing because the state-mandated writing test is handwritten.
- My school's (district's) policy forbids using computers when teaching writing because it does not match the format of the state-mandated writing test.

As Russell and Abrams (in press) describe, while many states have reported increases in students' scores on state tests (e.g., Texas, Massachusetts, and California among others have all celebrated gains over the past few years), for some students these gains come at the expense of opportunities in school to develop skills in using computers, particularly for writing. Although the majority of teachers report that they do not believe that the use of computers for writing

has been affected by the testing program, 30.2% of teachers across the nation do believe that they are not using computers for writing because the state-mandated test is handwritten. Across the nation, a higher percentage of teachers in urban locations and in lower-performing schools, as compared to suburban and high-performing schools, believe they have decreased instructional use of computers for writing because of the format of the state test. Moreover, teachers in urban and lower-performing schools as compared to teachers in suburban and high-performing schools report that fewer of their students have access to computers at home or write regularly using computers. Thus, the same students whose teachers are more likely not to have them use computers for writing because of the format of the state test are significantly less likely to have computers in their homes and therefore are less able to develop proficiency in writing with computers (or working with computers more generally) outside of school.

Briefly, then, research on the mode of administration effect for writing tests highlights two aspects of validity that may be threatened by administering tests on paper or on computer. First, depending upon the prior experiences of students and the mode of test administration, the degree to which a test is able to measure the desired construct can be threatened by differences in students' ability to work in the given mode. Second, the mode in which the test is administered may also influence the instructional practices of teachers. In instances where this influence steers teachers away from effective instructional practices, namely, developing students' computer and computer-writing skills, there are negative consequences. Positive consequences, however, result when the influence encourages teachers to adopt or increase their use of effective instructional practices.

To reduce the mode of administration effect and to promote instructional uses of computers for writing, Russell and his colleagues have suggested that students be given the option of composing written responses for state-level tests on paper or with a word processor. As noted above, state educational officials, however, have raised concerns that this policy might place students in urban and/or under-funded districts at a disadvantage because they might not have computers available to them in school if they were given the option of using them during testing. The data presented above, however, suggest that these students are already being placed at a disadvantage because their teachers are more likely to discourage them from using computers for writing in school. Additionally, their schools are nearly three times as likely to have policies that prohibit use of computers for writing. Moreover, as noted previously, these same students are significantly less likely to have computers in their homes and therefore have limited opportunities to master even the most basic computer skills (i.e., proficiency in keyboarding and facility in writing with computers). Despite rising test scores, teachers and school policies that decrease the use of computers coupled with limited access to computers at home is under-preparing many students in urban and poorly performing schools for the workplace. While these policies may help increase the performance of students on state

tests, the tradeoff between improved test scores and increased ability to work with computers may prove expensive as these students enter the workforce.

In addition, since the mode of administration effect reported by Russell and his colleagues only occurs for students who are accustomed to writing with computers, and because students in suburban and high-performing schools are much more likely to be accustomed to writing with computers, state tests are likely under-representing the difference in academic achievement between urban and suburban schools. Despite concerns about the need to close the gap between urban and suburban schools, the current policy which prohibits use of computers during state tests calls into question the use of these tests to examine the achievement gap.

Elementary and Secondary School Testing Policies and Validity of Computer-Based Tests

Much of the research on computer-based testing has approached the issue of validity by examining the extent to which scores provided by computer-based tests are comparable to scores provided by their paper-based predecessors. In some cases, cross-modal score comparability has been examined for sub-populations, with a specific focus on whether and how score comparability varies with prior computer experience. Without question, this body of research has been invaluable in advancing the quality of computer-based tests. It has highlighted the importance of carefully planning item layout, the need to provide examinees with the ability to review and revise responses, challenges in using text that spans more than one screen, as well as challenges to presenting mathematics problems that require examinees to solve problems using scratch space. In addition, this research shows that for some content areas and test formats, prior computer experience is an important factor that effects the validity of scores provided by computer-based tests. Despite these challenges, much of this research provides evidence that well designed computer-based tests can provide valid information about examinees performance in a wide variety of content domains.

It is important to note, however, that the vast majority of this research has focused on adults rather than elementary and secondary students. As required by *No Child Left Behind Act of 2001* (Public Law No: 107-110), state-level testing programs are required to test students in Grades 3–8. While findings from studies conducted with adult populations may generalize to students in elementary and middle schools, little evidence currently exists to support this assumption. Similarly, as Kolen and Brennan (1995) argue, factors that affect the validity of computer-based tests are likely test specific. Thus, to establish the validity of scores provided by computer-based tests employed by state-level testing programs, test developers cannot rely on studies conducted on different tests administered to different populations. Instead, validity studies should be conducted as each test is transitioned from paper to computer or when a new

computer-based test is introduced.

As these studies are conducted, it is important to consider the ways in which students learn and produce work in the classroom and the ways in which they are able to produce responses on the test. As research on computers and writing demonstrates, discrepancies between the way in which students produce writing while learning and the way in which they produce writing during testing has a significant impact on the validity of information provided by writing tests. Similarly, research on some mathematics tests indicates that validity is threatened when students experience difficulty accessing scratch space in which they perform calculations or produce diagrams while solving a given problem. It is unclear whether similar issues may exist for other item formats or in other content areas.

Finally, it is important to note that the majority of research on the validity of information provided by computer-based tests has focused on norm-referenced exams for which an individual's test score is compared to scores for a larger body of examinees. Most state-level testing programs, however, do not employ norm-referenced tests. Instead, as is required by *No Child Left Behind Act of 2001* (Public Law No: 107-110), state-level tests are criterion-referenced, in which each student's performance is compared to a pre-defined standard. For norm-referenced tests, test developers have traditionally emphasized the need to create standardized conditions in which all examinees perform the test. The need for standardized conditions is important so that direct comparisons between students can be made with confidence that any differences between examinees' scores result from differences in their ability or achievement rather than differences in the conditions in which they took the test. As norm-referenced tests transition from paper-based to computer-based administration cross-modal comparability is important so that scores provided by either mode can be compared with each other.

For standards-based exams, however, the need for standardized conditions may not be as necessary. While this notion departs from traditional beliefs, it is important to remember that the purpose of most state-level standards-based tests is to determine whether each individual has developed the skills and knowledge to an acceptable level. Thus, each examinee's test performance is compared to a standard rather than to the performance of all other examinees. In addition, the decisions made based on an examinee's test performance focus on the extent to which the examinee has developed the skills and knowledge defined as necessary to meet a given performance standard. Given that decisions are made based on an examinee's test performance, each examinee should be provided with an opportunity to demonstrate the skills and knowledge that they have developed. As an example, if an examinee is best able to demonstrate their writing skills on paper, then that examinee should be allowed to perform the test on paper. Similarly, if an examinee performs better by using a computer, then the examinee should be provided access to a computer while being tested. Moreover, if the ability of examinees to demonstrate their best performance is affected by factors such as the size of fonts, the type of scratch space

provided, or the type of calculator used (provided the test does not measure basic arithmetic skills), then examinees should be able to select or use those tools that will allow them to demonstrate their best performance.

In essence, the notion of allowing each examinee to customize their testing experience such that they are able to demonstrate their best performance is consistent with the practice of providing accommodations to students with special needs. To be clear, we are not advocating that examinees be provided with access to any tool or test format such that the accommodation itself leads to a higher test score. Rather, we suggest that examinees should be able to customize the environment in which they perform the test so that the influence of factors irrelevant to the construct being measured is reduced. Furthermore, when making decisions about the type of customizations that are allowed, we suggest that the standards used to determine whether or not to allow an accommodation for students with special needs be applied. As Thurlow et al. (2000) describe, before any accommodation is allowed three conditions must be met. First, it must be established that the accommodation has a positive impact on the performance of students diagnosed with the target disability(s) – or in the case of customization, those students who believe they will benefit by the customized feature. Second, the accommodation should have no impact on the performance of students that have not been diagnosed with the target disability (that is, providing the accommodation does not provide an unfair advantage). And third, the accommodation does not alter the underlying psychometric properties of the measurement scale.

While we recognize that the idea of allowing examinees to customize the test conditions will not sit well with many readers, we believe such a practice has important implications for efforts to examine the validity of scores provided by standards-based tests administered on computers. First, by acknowledging that sub-sets of examinees will perform differently on the test depending upon a given condition or set of conditions, there is no longer a need to focus on cross-modal or cross-condition comparability. Second, instead of identifying sub-populations of students that may be adversely affected by a test administered on computer, the focus shifts to identifying sub-populations for whom more valid measures of their achievement would be obtained under a given condition or set of conditions. Third, studies would be undertaken to ensure that a given condition did not artificially inflate the performance of a sub-population of students. Finally, when deciding whether to allow a given condition to be modified for a sub-set of examinees, greater emphasis would be placed on examining the extent to which test scores provided under the condition provide information that is more consistent with the performance of the examinees in the classroom – in essence increasing the importance of collecting information to examine concurrent validity.

Without question, decreasing emphasis on standardization and increasing emphasis on customized conditions would challenge efforts to transition testing to a computer-based format. However, given the important and high-stakes decisions made about students and schools based on standards-based

tests, it is vital that these tests provide accurate and valid estimates of each student's achievement. Just as accommodations have been demonstrated to increase the validity of information provided by tests for students with special needs, similar increases in validity could result by providing examinees with more flexibility in customizing the conditions under which they are asked to demonstrate their achievement. Already, there is evidence in the area of writing that a standardized condition, whether it be paper- or computer-based, adversely affects the validity of scores for sub-groups of students. Thus, in the area of writing, validity would be increased by providing examinees with the flexibility to choose the condition in which they can best demonstrate their writing skills. As the viability of administering standards-based tests on computer continues to increase, similar research and flexibility is needed in all areas of standards-based testing.

Today, state-level testing programs have reached the watershed of computer-based testing. While twelve states have already begun actively exploring the transition to computer-based delivery, over the next few years, several other states are also likely to reach this transition point. Without question, computer-based testing holds promise to increase the efficiency of testing, but it also has potential to increase the validity of information provided by the standards-based tests. To do so, however, current emphasis on cross-modal comparability and standardization of administration conditions may need to be relaxed so that examinees can have more control over the construct-irrelevant factors that may interfere with their performance in the tested domain.



References

- 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.
- Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, 10(18). Retrieved February 1, 2003 from: <http://epaa.asu.edu/epaa/v10n18.html>
- Becker, H. J. (1999). *Internet use by teachers: Conditions of professional use and teacher-directed student use. Teaching, learning, and computing: 1998 national survey* (Report No. 1). Irvine, CA: Center for Research on Information Technology and Organizations.
- Biskin, B. H., & Kolotkin, R. L. (1977). Effects of computerized administration on scores on the Minnesota Multiphasic Personality Inventory. *Applied Psychological Measurement*, 1(4), 543–599.
- Bresloin, M. J., Jr. (1984). A comparative study of computer administration of the Minnesota Multiphasic Personality Inventory in an inpatient psychiatric setting. Unpublished doctoral dissertation, Loyola University, Chicago, IL.
- Elwood, D. L. (1969). Automation of psychological testing. *American Psychologist*, 24(3), 287–289.
- Elwood, D. L., & Griffin, H. R. (1972). Individual intelligence testing without the examiner: Reliability of an automated method. *Journal of Consulting and Clinical Psychology*, 38(1), 9–14.
- Evan, W. M., & Miller, J. R. (1969). Differential effects on response bias of computer vs. conventional administration of a social science questionnaire: An exploratory methodological experiment, *Behavioral Science*, 14(3), 216–227.
- Hedl, J. J., Jr., O'Neil, H. F., & Hansen, D. H. (1973). Affective reactions toward computer-based intelligence testing. *Journal of Consulting and Clinical Psychology*, 40(2), 217–222.
- Hitti, F. J., Riffer, R. L., & Stuckless, E. R. (1971). *Computer-managed testing: A feasibility study with deaf students*. Rochester, NY: National Technical Institute for the Deaf.
- Johnson, D. E., & Mihal, W. L. (1973). Performance of Blacks and Whites in computerized versus manual testing environments. *American Psychologist*, 28(8), 694–699.
- Johnson, D. E., & White, C. B. (1980). Effects of training on computerized test performance on the elderly. *Journal of Applied Psychology*, 65(3), 357–358.

- Katz, L., & Dalby, T. J. (1981a). Computer and manual administration of the Eysenck Personality Inventory. *Journal of Clinical Psychology, 37*(3), 586–588.
- Katz, L., & Dalby, T. J. (1981b). Computer assisted and traditional psychological assessment of elementary-school-aged-children. *Contemporary Educational Psychology, 6*(4), 314–322.
- Kolen, M. J. & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.
- Koson, D., Kitchen, C., Kochen, M., & Stodolosky, D. (1970). Psychological testing by computer: Effect on response bias. *Educational and Psychological Measurement, 30*(4), 803–10.
- Lee, J. A., & Hopkins, L. (1985, March). *The effects of training on computerized aptitude test performance and anxiety*. Paper presented at the annual meeting of the Eastern Psychological Association. Boston, MA.
- Lee, J. A., Moreno, K. E., & Sympson, J. B. (1986). The effects of test administration on test performance. *Educational and Psychological Measurement, 46*(2).
- Llabre, M. M., Clements, N. E., Fitzhugh, K. B., & Lancelotta, G. (1987). The effect of computer-administered testing on test anxiety and performance. *Journal of Educational Computing Research, 3*(4), 429–433.
- Lushene, R. E., O'Neil, H. F., & Dunn, T. (1974). Equivalent validity of a completely computerized MMPI. *Journal of Personality Assessment, 38*(4), 353–361.
- Mazzeo, J., Druesne, B., Raffeld, P., Checketts, K. T., & Muhlstein, E. (1991). *Comparability of computer and paper-and-pencil scores for two CLEP general examinations*. (ETS Report No. RR-92-14). Princeton, NJ: Educational Testing Service.
- Mazzeo, J. & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature* (ETS Report No. RR-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*(4), 325-33.
- Meyer, L., Orlofsky, G. F., Skinner, R. A., & Spicer, S. (2002). The state of the states. Quality Counts 2002: Building blocks for success: State efforts in early childhood education. *Education Week, 21*(17), 68–170.
- Mueller, D. J., & Wasser, V. (1977). Implications of changing answers on objective test items. *Journal of Educational Measurement, 14*(1), 9–14.

- Pedulla, J., Abrams, L., Madaus, G., Russell, M., Ramos, M., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Boston: Boston College, National Board on Educational Testing and Public Policy. Retrieved March 10, 2003 from <http://www.bc.edu/research/nbetpp/statements/nbr2.pdf>
- Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*, 7(20). Retrieved February 1, 2003 from <http://epaa.asu.edu/epaa/v7n20/>
- Russell, M., & Abrams, L. (in press). Instructional uses of computers for writing: How some teachers alter instructional practices in response to state testing. *Teachers College Record*.
- Russell, M., O'Brien, E., Bebell, D., & O'Dwyer, L. (2003). *Students' beliefs, access, and use of computers in school and at home*. Boston: Boston College, Technology, and Assessment Study Collaborative.

