Research Memorandum

Online Assessment and the Comparability of Score Meaning

Randy Elliott Bennett

Research & Development

November 2003
RM-03-05

Online Assessment and the Comparability of Score Meaning

Randy Elliott Bennett

ETS, Princeton, NJ

November 2003

Abstract

Comparability refers to the commonality of score meaning across testing conditions including delivery modes, computer platforms, and scoring presentation. As state education agencies introduce online tests and online scoring of constructed-response questions, score comparability becomes important. For example, it should be a matter of indifference to the examinee whether the test is administered on computer or paper, or whether it is taken on a large-screen display or a small one. To the extent that test results are invariant across such conditions, scores may be used interchangeably. This paper explores the comparability issues related to online assessment at the elementary and secondary level, and reviews the available research.

Key words: Comparability, online tests, online scoring, computer-based testing, display, font size, presentation, resolution, screen size

Acknowledgements

The positions expressed in this paper are those of the author and not necessarily of ETS.

I thank Manfred Steffen, Tim Davey, and Dan Eignor for their comments on an earlier draft of this paper.

Online Assessment and the Comparability of Score Meaning

At least a dozen U.S. states are moving components of their assessment systems to online delivery. These states include Arkansas, Georgia, Idaho, Kansas, Kentucky, Maryland, North Carolina, Oregon, South Dakota, Texas, Utah, and Virginia (Bennett, 2002; Olson, 2003). Seven more—Delaware, Illinois, Indiana, Mississippi, New York, Pennsylvania, and Washington—and the District of Columbia have run or are running pilots to decide whether to start the transition.

If one looks across these state efforts, several things are apparent. First, the efforts are being pursued at both the elementary and secondary levels, in all key content areas (reading, math, science, and social studies), and for regular and special education populations. Second, the efforts involve low- and high-stakes assessments: diagnostic and formative tests, progress measures, and promotion and graduation exams. Third, the activities vary widely in their progress and target implementation dates. The earliest adopters already have administered significant numbers of tests online: through June 2003, 135,000 in North Carolina, 165,000 in Virginia, and more than a million classroom tests in Georgia (Bazemore, 2003; Harmon, 2003; Loving-Ryder, 2003). Fourth, these efforts initially use multiple-choice items almost exclusively, not only because the software for doing that is more evolved but also because moving a testing program to computer is a complicated process that would only be made more difficult by including constructed-response questions in the initial phases. Finally, several of the state efforts, like those in Virginia and South Dakota, are part of an explicit plan to use technology for broad educational purposes that go well beyond assessment.

Why are states moving to online assessment? There are several reasons but chief among them is that scoring and reporting can be done more rapidly. A second reason is that the assessment can be adapted to individual student characteristics. Third, the costs of testing may eventually be reduced. Finally, there is the promise of being able to measure skills on computer that cannot be assessed on paper, such as using technology for problem solving (Bennett, Jenkins, Persky, & Weiss, 2003; Bennett & Persky, 2002).

Although online assessment is attractive, those states attempting it have encountered significant challenges that, in some cases, have delayed implementation efforts considerably. These challenges include the up-front costs of equipment, connectivity, staff training, delivery software, and item banking; the tight timelines established by the federal *No Child Left Behind* legislation for putting comprehensive assessment programs into place; the lack of school staff

available to keep equipment running properly; the security threats to electronically delivered tests, especially when all students are not tested at the same time; and issues related to measurement and fairness. This paper focuses on measurement and fairness issues but, in particular, the comparability of score meaning with respect to delivery modes, computer platforms, and scoring presentation.

What Is Comparability and When Is It Important?

For purposes of this paper, comparability refers to the commonality of score meaning across testing conditions including delivery modes, computer platforms, and scoring presentation. When comparability exists, scores from different testing conditions can be used interchangeably. For example, scores derived from one condition can be referenced to norms collected, or to cut scores set, in another.

The American Psychological Association's *Guidelines for Computer-based Tests and Interpretations* (American Psychological Association [APA], 1986, p. 18) states that scores may be considered equivalent when individuals are rank ordered in approximately the same way and when the score distributions are approximately the same. If the rank-ordering criterion is met but the distributions are not the same, it may be possible to make scores interchangeable by equating. Although the *Guidelines* pose these criteria in the context of comparability across computer and paper delivery, the criteria are generally applicable to any difference in testing conditions between instruments intended to reflect the same construct.

Comparability is required when scores need to have common meaning with respect to one another, to some reference group, or to a content standard. If scores are not comparable across delivery modes, scoring presentation, or computer platforms, and the test varies along one or more of these dimensions, the decisions we make from assessment may be wrong. Wrong decisions may be made about individuals for such things as promotion, graduation, diagnosis, or progress reporting. Wrong decisions also may be made about institutions. For example, under *No Child Left Behind*, schools could under- or overestimate their standing with respect to Adequate Yearly Progress, the federal law's performance metric, because the conditions under which they test differ. Similarly, national assessments, such as the National Assessment of Educational Progress (NAEP), could incorrectly estimate what the nation's school children know and can do.

Finally, these wrong decisions may unfairly impact population groups when the lack of comparability is associated more with some types of individuals or institutions than others.

Comparability of Delivery Modes

For state and national assessments, comparability across delivery modes is important because such assessments will almost certainly need to be delivered on both computer and paper, at least for the near-term. This need for dual delivery exists because most schools don't yet have enough computers to test all students in a particular cohort simultaneously and because some students still don't have the skills needed to respond to a test on computer effectively. At some point, schools will have the requisite hardware, students the needed skill, and computer tests the measurement and efficiency advantages that paper cannot match. But until that time, comparability of delivery modes will remain an issue.

The scores from paper and computer versions of the same test can diverge for several reasons. For one, differences in such *presentation characteristics* as the number of items on the screen versus the number on the printed page, or the size of text fonts, could impact performance. Differences in *response requirements* also may affect scores. For example, to take a paper test, a student need only know how to use a pencil to mark correct answers to multiple-choice problems and how to write in answers for the open-ended questions. In contrast, a computer-based test may require additional skills. The examinee may have to point, click, drag, drop, and scroll with the mouse, as well as use the keyboard to enter and edit text. A third reason scores could diverge across delivery modes is differences in *general administration characteristics*. For instance, the online test might present items adaptively, so that every examinee gets a different test, while the paper test contains the same items in the same order for all students. Additionally, the paper administration would typically require students to wait until time elapses before moving on to the next section, while the online administration may permit them to proceed whenever they are ready. Finally, the computer test may unintentionally be more speeded than the paper version, as it often takes longer to read and respond to text-laden questions on screen.

Many studies have investigated the comparability of paper and computer tests among adults. Mead and Drasgow (1993) reported on a meta-analysis of studies that estimated the correlation between testing modes after correcting for unreliability and compared differences in

mean performance across modes. Based on 159 estimates, they found the correlation for timed power tests like those used in educational settings to be .97, suggesting score equivalence, but the correlation for speeded measures, such as clerical tests, to be .72. For the timed power tests, the standardized mean difference between modes was .03 and the standard deviation of the differences was .15. Computerized tests, therefore, were harder than paper versions, but only trivially so, and the variation in this mode difference from one study to the next was minimal.

Gallagher, Bridgeman, and Cahalan (2000) examined comparability for population groups on the Graduate Record Examinations® (GRE®) General Test, Graduate Management Admission Test® (GMAT®), SAT® I: Reasoning Test, Praxis: Professional Assessment for Beginning Teachers®, and Test of English as a Foreign Language™ (TOEFL®). These investigators discovered that delivery mode consistently changed the size of the differences between focal and reference group performance for some groups, but only by small amounts. For African American and Hispanic students, for example, the difference in performance relative to White students was smaller on computer-based tests than on paper tests. From one mode to the other, the difference in performance between groups changed by up to .25 standard deviation units, depending upon the test. For White females, the difference relative to White males was smaller on the paper versions than on the online editions. This difference changed as a function of delivery mode by up to .14 standard deviations, again depending upon the particular test.

At the elementary and secondary school level, the data are far more limited simply because of the novelty of computer-based testing for this population. Among the studies with large samples are those sponsored by the Oregon Department of Education and the North Carolina Department of Public Instruction. Choi and Tinkler (2002) assessed some 800 Oregon students in each of third and tenth grades with multiple-choice reading and mathematics items delivered on paper and by computer. They discovered that items presented on computer were generally harder than items presented on paper, but that this difference was more apparent for third graders and for reading than for math tests. For the North Carolina Department of Public Instruction, Coon, McLeod, and Thissen (2002) evaluated third graders in reading and fifth graders in math, with roughly 1,300 students in each grade taking paper test forms and 400 students taking the same test forms on computer. All items were multiple-choice. Results indicated that for both grades, scores were not comparable, with scale scores being higher for paper than for the online tests. Further, within grades, the mode differences were not the same

across forms and, for one form, there was a significant delivery-mode by ethnic-group interaction, indicating the possibility that mode differences varied among population groups. This lack of consistency suggests that comparability could not be achieved for these particular tests by a simple equating of scores from mode to the other using data from the total student group.

Given the lack of comparability apparent in the few studies conducted with multiple-choice tests, one might expect similar results from tests composed of constructed-response questions. Here, the threat to comparability should be greater because the online version will make heavier demands on computer skill than an electronic test consisting solely of multiple-choice questions. These greater demands could increase the chances for an interaction with computer proficiency, such that students who routinely do academic work on the computer are more accurately assessed in that mode while others may be better tested on paper. The research on this question, too, is very limited because most states have not yet included constructed-response items in their online assessments. But the few (relatively small) studies which have been done suggest that scores from productive writing tests, for example, may not be the same across delivery mode and that computer experience may interact with mode in determining performance (e.g., Russell & Haney, 1997; Russell, 1999; Russell & Plati, 2001; Wolfe, Bolton, Feltovich, & Niday, 1996).

Comparability of Computer Platforms

One of the attractions of Internet testing is that, in principle, any connected computer can run the test. This fact means that tests can be delivered to a wide variety of locations, including school computer labs and classrooms. The hardware and software configurations in these locations, however, will undoubtedly differ. This variation may have measurement consequences because the presentation of items may not be the same from one machine to the next.

One way in which item presentation can be affected is by the Internet connection. This connection is, in reality, not a single electronic link but a chain of connections between the test center and the testing agency. In test delivery models that fetch questions from a remote server one at a time, the flow of information through this chain dictates how long an examinee will need to wait before seeing the next item. That delay will be determined by several factors. The first factor is the test center link to its Internet service provider (ISP). The quality of this connection is determined by two things: its bandwidth and the number of computers that are actively sharing it.

For example, if the test center is in a school district, all of the district's classroom and administrative computers may go to the ISP through the same high-speed line, which typically will effectively support only a portion of those computers simultaneously. This arrangement means that response time on test center machines may be slowed if many other computers in the district are accessing the Internet during the testing session. The chain between the test center and testing agency also will be affected by conditions at the ISP itself. Individual ISPs do occasionally encounter problems and, when they do, all traffic entering or exiting the Internet through them may come to a halt. Third, the Internet has an impact. If demand is high because of time of day, or an unusual news event, response time everywhere may slow. Quality may be affected too by the testing agency's ISP and, of course, by the testing agency server itself.

Besides the Internet connection, item presentation may be influenced by other factors, including differences in screen size, screen resolution, operating system settings, and browser settings (Bridgeman, Lennon, & Jackenthal, 2001).

How does screen size impact item presentation? All other things equal, differences in screen size do *not* affect the amount of information displayed. Smaller monitors make the same information look smaller because the text displayed on such monitors is, in fact, physically littler. As a result, a question presented on a smaller monitor may be harder to read than the same question presented on a larger one. But the amount of information displayed will be the same on both screens.

What about resolution? Resolution affects the size of text *and* may affect how much information is shown. Given the same screen size and font size, text displayed at high resolution will be smaller than text displayed at a lower resolution. The higher-resolution screen is packing more pixels (picture elements) into the same physical area, so the pixels themselves will be smaller. As a result, a text character containing a fixed number of pixels will be smaller on the higher- than on the lower-resolution monitor. And because the text is smaller, there can be more of it. Higher resolution allows more words per line and lines per screen.

Figure 1 shows a reading comprehension item from Bridgeman et al. (2001) displayed in high resolution (1024 by 768). Notice that the entire passage fits on the screen and that each option takes only one line. In particular, note that the first line of the passage contains two complete sentences. Figure 2 depicts the same item in a much lower resolution (640 by 480). Notice that the text lines break differently—now only the first sentence fits on the opening line.

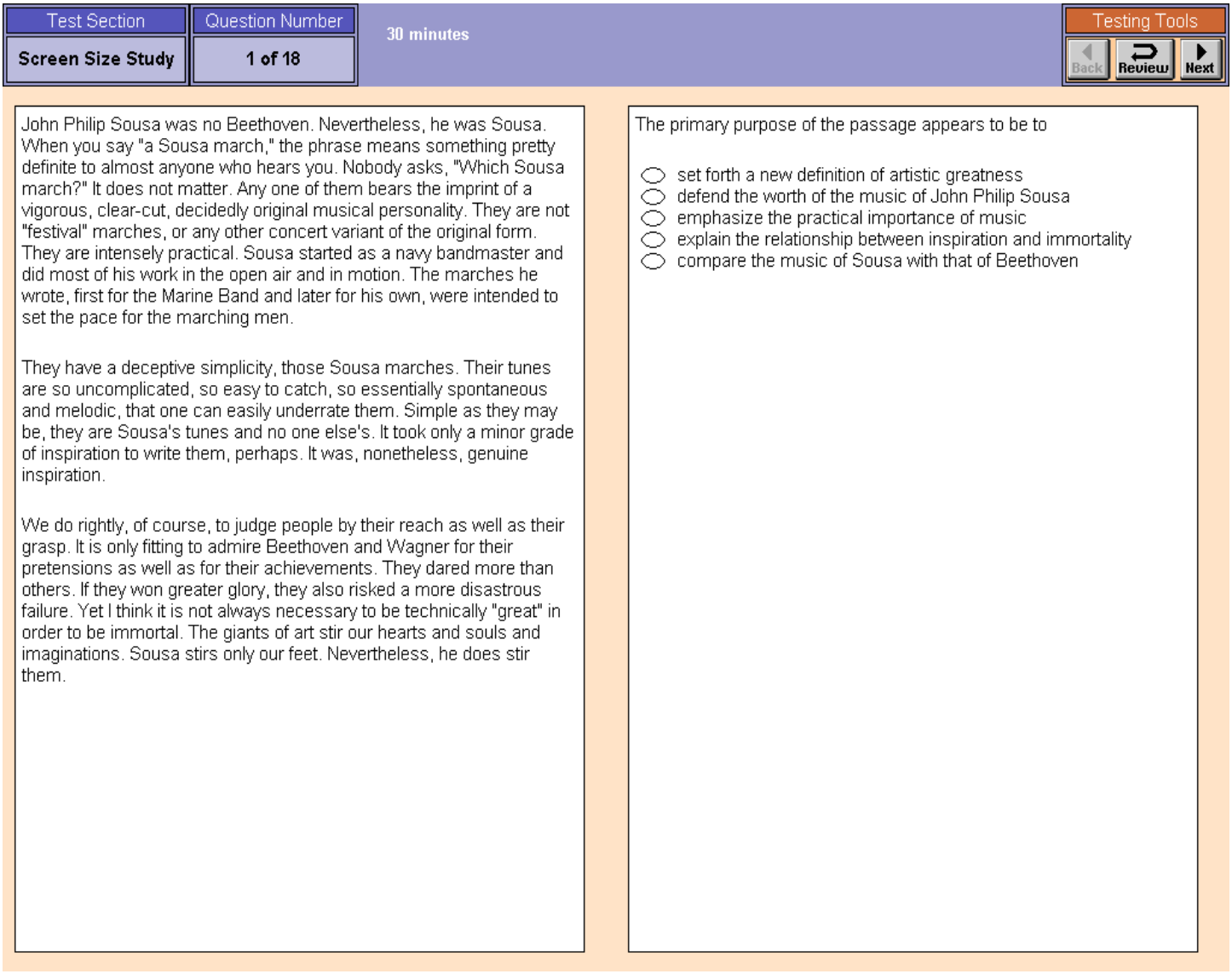

Figure 1. **A reading comprehension item presented in high resolution (1024 by 768).**

Note. From *Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance* (RR-01-23) (p. 3) by B. Bridgeman, M. L. Lennon, & A. Jackenthal, 2001. Princeton, NJ: ETS.

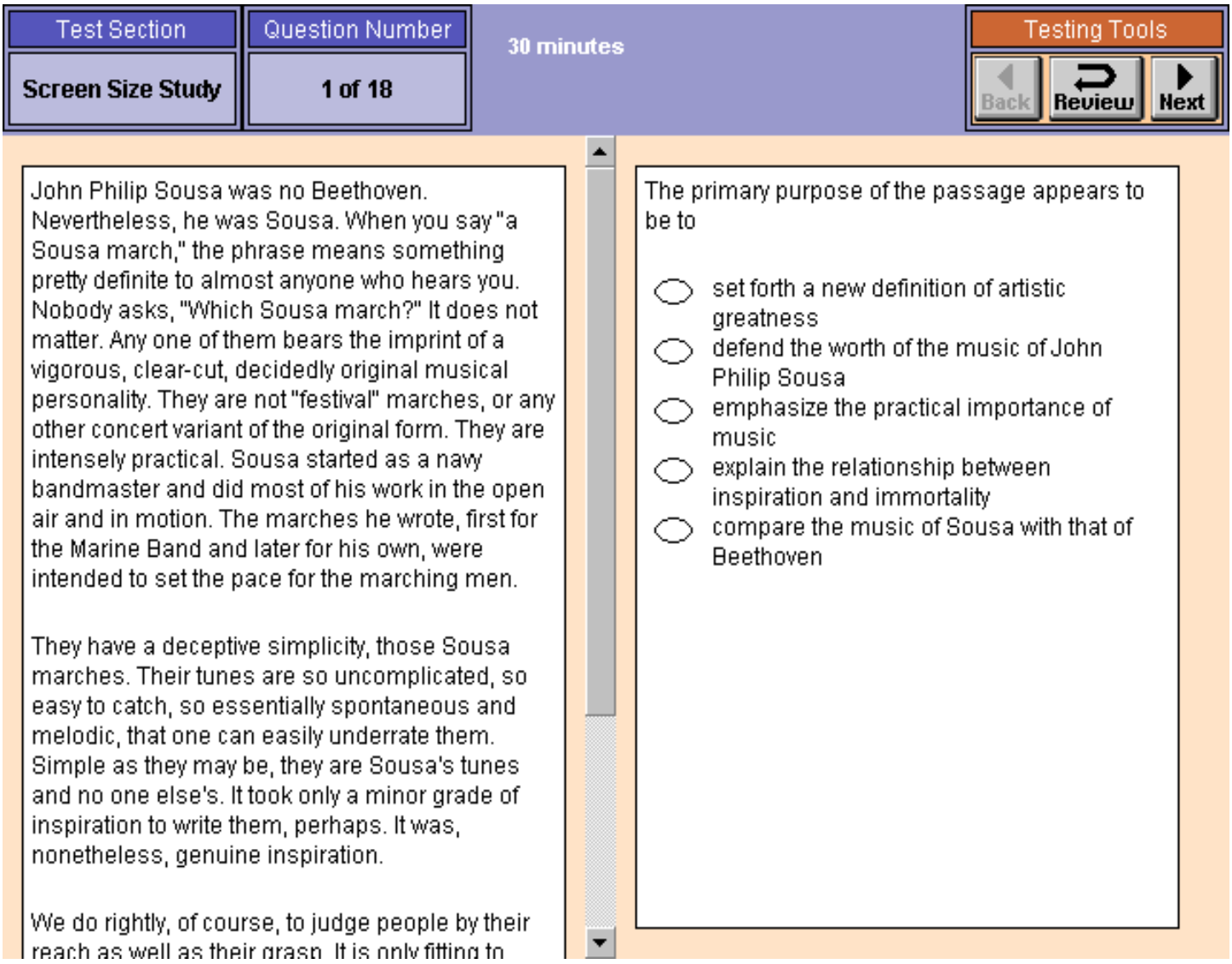

Figure 2. **A reading comprehension item presented in low resolution (640 by 480).**

Note. From *Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance* (RR-01-23) (p. 3) by B. Bridgeman, M. L. Lennon, & A. Jackenthal, 2001. Princeton, NJ: ETS.

Also note that the examinee must scroll to read the complete passage. Finally, the answer options each take more than one line.

The practical impact of these differences in screen resolution, then, is that lower resolutions may require examinees to spend more time locating information. Why? Because, they may need to do more scrolling and make more visual sweeps to process shorter lines of text. Lower resolutions also may increase processing difficulty if critical information is split between screens, perhaps making more prominent than intended the role of short-term memory in item solution.

What about font size? What do differences in font size imply? Font size affects the size of text and the amount of information displayed. That is, smaller letters allow the display of more information. If resolution is held constant, smaller fonts permit more characters per line and lines per screen. Unfortunately, font size can be changed in multiple ways, including through operating system settings, browser settings, and Web-page coding. As a result, font size may not be identical across machines or test centers under some Internet delivery models.

Figure 3 shows a reading comprehension item from Bridgeman et al. (2001) displayed in low resolution (640 by 480). The font size has been set to "small" in the Microsoft Windows control panel and "smallest" in the browser. The passage pane on the left side of the window shows one partial paragraph at the top followed by two complete paragraphs. The item stem and all five options are visible in the right-hand pane.

The same question is shown in Figure 4, this time displayed at the identical resolution but with the font size set to "large" in the Microsoft Windows control panel and "medium" in the browser. Now *only* the last paragraph in the passage is visible in the left pane. Also, the lines of text are shorter. Finally, only three of the five question options appear.

The effect of changing font size, then, is similar to the effect of allowing adjustments to resolution. At one extreme, the examinee may need to spend more time locating information or may have to do different cognitive processing. At the other, the text will be smaller and possibly harder to read.

What's the effect of differences in item presentation on test scores? Unfortunately, there appears to be almost no research that addresses this question directly and systematically in an assessment context. The most relevant study was conducted by Bridgeman, Lennon, and Jackenthal (2003). They looked at the effect of variations in screen size, resolution, and

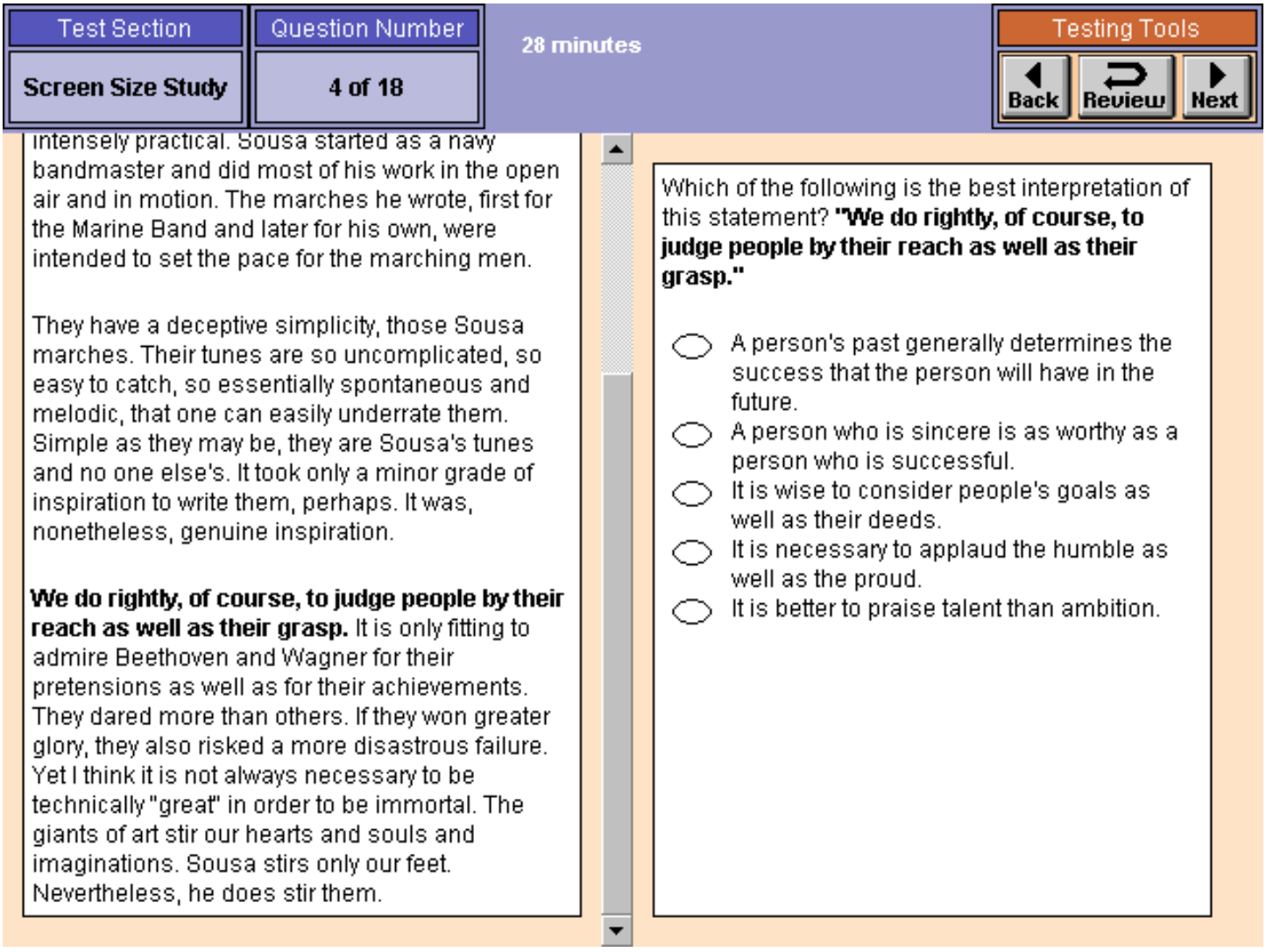

Figure 3. **A reading comprehension item presented with font size set to "Small" in the Microsoft Windows control panel and "Smallest" in the browser (640 by 480 resolution).**

Note. From *Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance* (RR-01-23) (p. 4) by B. Bridgeman, M. L. Lennon, & A. Jackenthal, 2001. Princeton, NJ: ETS.

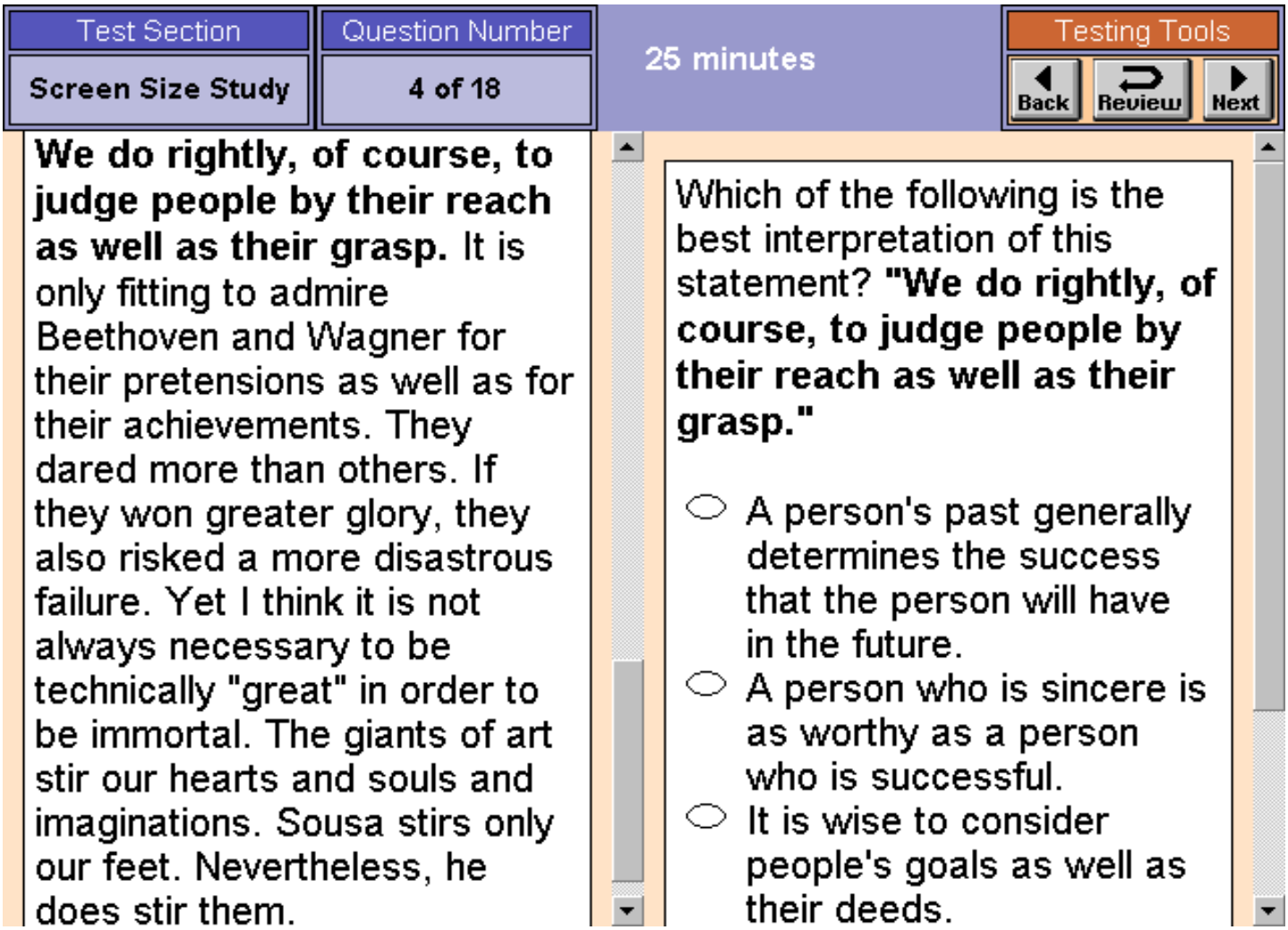

Figure 4. **A reading comprehension item presented with font size set to "Large" in the Microsoft Windows control panel and "Medium" in the browser (640 by 480 resolution).**

Note. From *Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance* (RR-01-23) (p. 4) by B. Bridgeman, M. L. Lennon, & A. Jackenthal, 2001. Princeton, NJ: ETS.

item-presentation latency on test performance for SAT I: Reasoning Test items. These investigators randomly assigned 357 high-school juniors to a variety of item presentation conditions. Two tests were administered, one consisting of quantitative comparisons questions and one of multiple-choice comprehension questions with associated reading passages. Bridgeman and his colleagues found no effect on math scores. Reading comprehension scores, however, were higher by about .25 standard deviations for students using a larger, higher-resolution display than for students using a smaller, lower resolution screen. (The effects of screen size and resolution could not be separated in the analysis.) Finally, the only test feature rated as interfering by the majority of students was scrolling. Bridgeman et al. (2001) suggest that a prudent approach in Web delivery of high-stakes tests would be to attempt to have comparable scrolling across computer configurations.

Fortunately, variation in item presentation can be controlled, at least to a substantial degree. One approach is to establish hardware and software standards to limit presentation differences. Delivery to a limited range of configurations is the approach that many high-stakes computer-based testing programs, such as the GRE General Test and GMAT, currently take. A second possibility is to manipulate presentation characteristics through choice of the Internet delivery model. That is, use testing software that adjusts font size, font type, and resolution by taking control of the examinee machine at the operating system level. This approach may not be the first choice of school test centers, however, because it typically requires installation of special software on center machines. Absent such fine control through software, a third possibility is to have proctors set display characteristics before starting the test and reset them after the examination concludes, so that other software used on those machines runs properly. A fourth possibility is to design items for the lowest common denominator, for example, for the lowest likely resolution to ensure that all answer options fit on a single screen. Finally, one can attempt to render items intelligently by automatically scaling text to wash out differences in resolution. If the resolution is high, make the text a little bigger, so that roughly the same information ends up on the screen as in the low-resolution display.

Comparability of Scoring Presentation

A final type of comparability associated with the use of technology relates to how constructed-responses are presented to human judges for scoring. One way in which this presentation may vary is whether responses—be they essays or answers to open-ended mathematics problem—are scored on paper or online. An advantage of the latter approach is that readers can be widely distributed, working from anywhere there is an Internet connection.

A second manner in which scoring presentation may vary is that responses may be submitted in either handwritten on typewritten form, depending upon the way the test was originally taken. Note that the question here is not how delivery mode affects the examinee's performance but, rather, how the form of the response affects the *scorer's judgment* of that performance. Also note that these two presentation conditions—scoring mode and format—are independent. Responses may be submitted in either handwritten or typewritten format. Handwritten responses can be presented for scoring as submitted or they can be scanned for computer display. Likewise, key-entered responses can be scored from printed copy or from a computer screen.

Online scoring has been employed for close to a decade, first by NCS Pearson in the NAEP, and more recently by ETS for the GRE General Test's Analytical Writing Assessment, the GMAT, Praxis™, and TOEFL. Surprisingly, there is very little published research in peer-reviewed journals on this type of comparability. The most comprehensive study is probably that conducted by Zhang, Powers, Wright, and Morgan (2003). These investigators studied scores from more than 11,000 high school students who had taken either the Advanced Placement Program® English Language and Composition test or the AP® Calculus test. All exams had been scored in handwritten, paper form through the operational AP reading by graders gathered in one or more common locations. As part of the reading, 500 of the exams for each subject were independently re-scored to estimate inter-rater reliability. Those 500 exams were subsequently scanned and again scored twice independently, this time by different readers working from individual home or office locations through the Internet. Zhang et al. compared the score level, variability, and inter-rater agreement for the individual questions composing each exam, as well as the overall exam passing rates, across modes. They found little, if any, practical difference on their comparability indicators (e.g., differences of less than .1 standard deviations for the English Language and Composition test and less than .03 standard deviations for the Calculus test).

These findings are basically consistent with an older study conducted for NAEP at grades 4, 8, and 12 in each of five subject areas (ETS, 1993), and with smaller studies using college-age students by Powers, Farnum, Grant, and Kubota (1997) for GMAT essays, and by Powers and Farnum (1997) for The Praxis Series: Professional Assessments for Beginning Teachers essays.

Whereas the limited available research has generally supported the comparability of online and traditional paper scoring, the same cannot be said for the scoring of handwritten versus typed responses. Powers, Fowles, Farnum, and Ramsey (1994) used answers from 32 college-age students, who each wrote two essays, one key-entered and one handwritten, in connection with pilot tests for the Praxis I Series: Academic Skills Assessments. Each essay was then transcribed to the other format, and the essays were presented on paper to two pairs of readers. The investigators found that the handwritten versions of the essays were graded significantly higher than the typed ones. In a second experiment, the investigators were able to reduce the scoring effect somewhat by training a new set of readers with both types of essays and calling their attention to the different impressions format might make. Powers and his colleagues concluded that the size of the performance difference was of little practical importance, in part because the essay was only one component of the Praxis I writing score (which also included performance on multiple-choice questions).

In a subsequent study, Powers and Farnum (1997), presented 40 Praxis essays on-screen and on paper, and in both typed and handwritten formats, to four pairs of readers. As noted above, the investigators found no differences for scoring on-screen versus on paper. However, they did find that the handwritten versions of essays were graded .2 standard deviations higher than the typed versions of the same text. This effect was virtually the same size as that found in the earlier study described above by Powers et al. (1994).

Russell and Tao (in press) replicated in the school population the scoring presentation effect that Powers and his colleagues found for college students. These investigators analyzed 52 essays in grade 4, and 60 in each of grades 8 and 10, written in response to questions taken from the Massachusetts Comprehensive Assessment System (MCAS) Language Arts Test. All essays were handwritten by students and subsequently typed on computer by the investigators. Once typed, the essays were presented on paper to six raters at each grade level in one of three formats—handwritten, printed in single spaced 12-point text, and printed in double-spaced 14-point text. The last condition was intended to correct for any difference in appearance between

handwritten and typed essays due to length. Russell found the handwritten versions to receive significantly higher scores than the typed ones, but detected no difference between typed essays of different apparent lengths.

As part of their study, Russell and Tao (in press) also asked two readers to identify and categorize errors in each of 30 grade 8 essays. Half of the essays appeared in handwritten form and half appeared in print, with the half seen in print by the first reader presented as handwritten to the second reader. The error categories were spelling, punctuation, capitalization, awkward transitions, and confusing phrases or sentences. Results showed that the two raters detected significantly more spelling errors and more confusing phrases or sentences when the essays were presented in print.

In a second study, Russell (in press) had handwritten eighth grade MCAS essays enter onto computer. He then, presented the responses on paper to eight readers in four formats—handwritten, printed in single-spaced type font, printed in single-spaced script font to simulate handwritten text, and printed in single-spaced font with all spelling errors corrected. Results showed that the scores for the handwritten versions and for the script-font versions did not differ from one another, but that both were graded significantly higher than the same essays represented in typed font. The effect on scores of corrected spelling was not significant for any comparison.

Russell (in press) next repeated the scoring with a second set of four readers trained to avoid the presentation effect. These readers graded only the original handwritten responses and their verbatim transcriptions in type font (i.e., without spelling corrected). In this scoring, the presentation effect was eliminated. That is, no significant difference was found between the scores awarded to the handwritten and typed versions.

In sum, the available research suggests little, if any, effect for computer versus paper display but a consistent difference for typed compared with handwritten presentation. Why typed essays receive lower scores is not completely clear, though the results of one study suggest that substantive and mechanical errors may stand out more since these responses are easier to read. However, other possibilities include that raters' expectations for typed responses are higher because such responses have the look of a final draft, and that typed text may encourage a weaker connection with the reader since the student's attempts at revision are not apparent.

Conclusion

Many states are moving components of their assessment systems to the computer or experimenting with that possibility. Although the promise of online assessment is substantial, states are encountering significant issues, including ones of measurement and fairness. Perhaps the most critical of these measurement and fairness issues relates to score comparability, in particular of delivery modes, computer platforms, and scoring presentation. Although some data at the elementary and secondary level are beginning to emerge, far more work needs to be done because variation in these conditions may affect scores in irrelevant ways that have undesirable consequences for institutions and for individuals. Particularly distressing is the potential for such variation to unfairly affect population groups, such as females, minority group members, or students attending schools in poor neighborhoods. School socio-economic status, for example, may be related to such conditions as quality of computer platform or computer familiarity that could, in turn, impact score comparability.

As noted, there is relatively little research on comparability issues at the elementary and secondary level. Further, what little research there is tends not to be published in peer-reviewed measurement journals but, rather, found in technical reports or in otherwise unpublished manuscripts. For online assessment to become a credible mechanism for making consequential decisions about students and institutions, the requisite research on comparability must not only be done, it must be reported in peer-reviewed scientific outlets where its quality can be vetted.

While the necessary research is being conducted, state assessment programs must try to control, or otherwise account for, variation known to impact scores. For differences in delivery mode, scores might be equated or, in the most extreme case, separate scales might be created for computer and paper versions. Equipment variation might be controlled by establishing hardware and software standards, directly manipulating font characteristics and resolution through the test delivery software, designing items so that they display adequately at the lowest likely resolution, or rendering items intelligently to limit the need for scrolling. Finally, when essay responses are handwritten by some examinees and typed by others, raters might be trained to avoid the scoring biases that appear to be associated with response format and not permitted to score operationally until they have demonstrated the ability to apply the same standards to both forms of response.

References

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations*. Washington, D.C.: Author.

Bazemore, M. (2003, June). NC online testing projects. In S. Lazer, S. Triplett, & R. Bennett (chairs), *Technology-based assessment: Lessons learned and future plans*. Pre-session conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, San Antonio, TX.

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment, 1*(1). Retreived August 4, 2003, from http://www.bc.edu/research/intasc/jtla/journal/v1n1.shtml

Bennett, R. E., Jenkins, F., Persky, H., & Weiss, A. (2003). Assessing complex problem-solving performances. *Assessment in Education, 10,* 347-359.

Bennett, R. E., & Persky, H. (2002). Problem solving in technology-rich environments. In Qualifications and Curriculum Authority (Ed.), *Assessing gifted and talented children* (pp. 19-33). London, England: Qualifications and Curriculum Authority.

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

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16,* 191-205.

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.

Coon, C., McLeod, L., & Thissen, D. (2002). *NCCATS update: Comparability results of paper and computer forms of the North Carolina End-of-Grade Tests* (RTI Project No. 08486.001). Raleigh, NC: North Carolina Department of Public Instruction.

ETS (1993). *The results of the NAEP 1993 field test for the 1994 National Assessment of Educational Progress*. Princeton, NJ: Author.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2000). *The effect of computer-based tests on racial/ethnic, gender, and language groups* (GRE Professional Board Report 96-21P, ETS RR-00-08). Princeton, NJ: ETS. Retrieved August 5, 2003, from ftp://ftp.ets.org/pub/gre/gre_96-21p.pdf

Harmon, D. J. (2003, June). Web-based testing: Georgia's five-year journey. In S. Lazer, S. Triplett, & R. Bennett (chairs), *Technology-based assessment: Lessons learned and future plans*. Pre-session conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, San Antonio, TX.

Loving-Ryder, S. (2003, June). Virginia Standards of Learning Web-based assessment. In S. Lazer, S. Triplett, & R. Bennett (chairs), *Technology-based assessment: Lessons learned and future plans*. Pre-session conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, San Antonio, TX.

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

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

Powers, D., & Farnum, M. (1997). *Effects of mode of presentation on essay scores* (ETS RM-97-08). Princeton, NJ: ETS.

Powers, D., Farnum, M., Grant, M., & Kubota, M. (1997). *A pilot test of online essay scoring*. Princeton, NJ: ETS.

Powers, D. E., Fowles, M. E., 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*, 220-233.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20). Retrieved August 5, 2003, from http://epaa.asu.edu/epaa/v7n20/

Russell, M., & Tao, W. (in press). Effects of handwriting and computer-print on composition scores: A follow-up to Powers et al. *Practical Assessment, Research and Evaluation*.

Russell, M. (in press). The influence of computer-print on rater scores. *Practical Assessment, Research and Evaluation*.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5*(3). Retrieved August 5, 2003, from http://epaa.asu.edu/epaa/v5n3.html

Russell, M., & Plati, T. (2001). Effects of computer versus paper administration of a state-mandated writing assessment. *TC Record.Org*. Retrieved April 19, 2002, from http://www.tcrecord.org/Content.asp?ContentID=10709

Wolfe, E. W., Bolton, S., Feltovich, B, & Niday, D. M. (1996). The influence of student experience with word processors on the quality of essays written for a direct writing assessment. *Assessing Writing, 3,* 123-147.

Zhang, Y. L., Powers, D. E., Wright, W., & Morgan, R. (2003). *Applying the online scoring network (OSN) to Advanced Placement Program® (AP®) tests* (ETS RR-03-12). Princeton, NJ: ETS.