Technology and Writing Assessment: Lessons Learned from the US National Assessment of Educational Progress[1]

Randy Elliot Bennett
Educational Testing Service
Princeton, NJ 08541
rbennett@ets.org

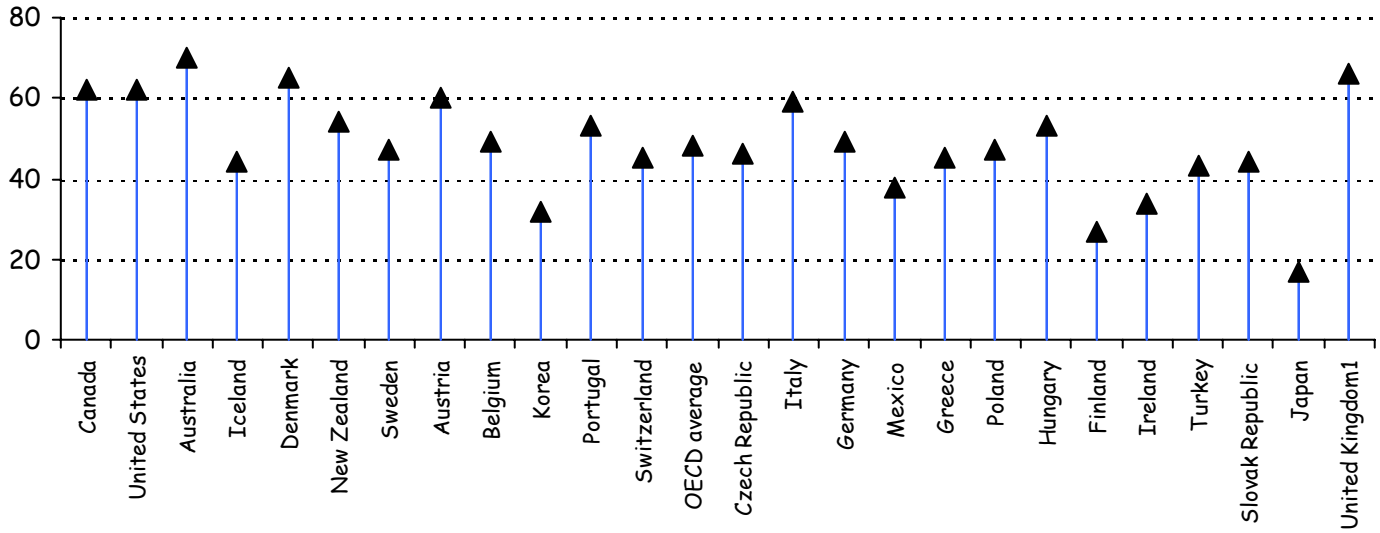
[1] This paper is from a keynote presentation made at the annual conference of the International Association for Educational Assessment, Singapore, May 2006. Portions of the paper were adapted from a previous presentation to the annual conference of the Dutch Examinations Association (Nederlandse Vereniging voor Examens [NVE]), Almelo, the Netherlands, November 2005.

Technology and Writing Assessment: Lessons Learned from the
US National Assessment of Educational Progress

In this paper, I review the results of two research studies conducted for the US National Assessment of Educational Progress (NAEP). NAEP is a sample survey of what US students in grades 4, 8, and 12 know and can do. The studies I review concern doing writing assessment on computer.

Why would one want to do writing assessment on computer? The answer is simple. The tools of choice in the workplace and in advanced academic environments have changed. The writing tool of choice is now the computer. At the school level, a similar transition is occurring, not only in the US, but in many other countries. PISA reports that among its participating countries, as of 2003, 48% of 15-year olds indicated using a word processor at least a few times a week (see Figure 1). Three years later, that percentage is surely higher.

Figure 1: Percentage of Students Reporting They Use Word Processing
"Almost every day" or "A few times each week" as of 2003



Note. From Schleicher (2006).

As students change the medium in which they routinely write, the idea of testing them in a different medium becomes problematic from validity, fairness, and credibility perspectives. Even so, for practical reasons the transition of large-scale writing assessment from paper to computer delivery will be a gradual one. Given that fact, some students may take their writing tests on paper while others take their tests on computer. Does it matter? That is, are the scores from paper and computer-based writing assessments comparable?

Are the Scores from Computer and Paper Writing Assessments Comparable?

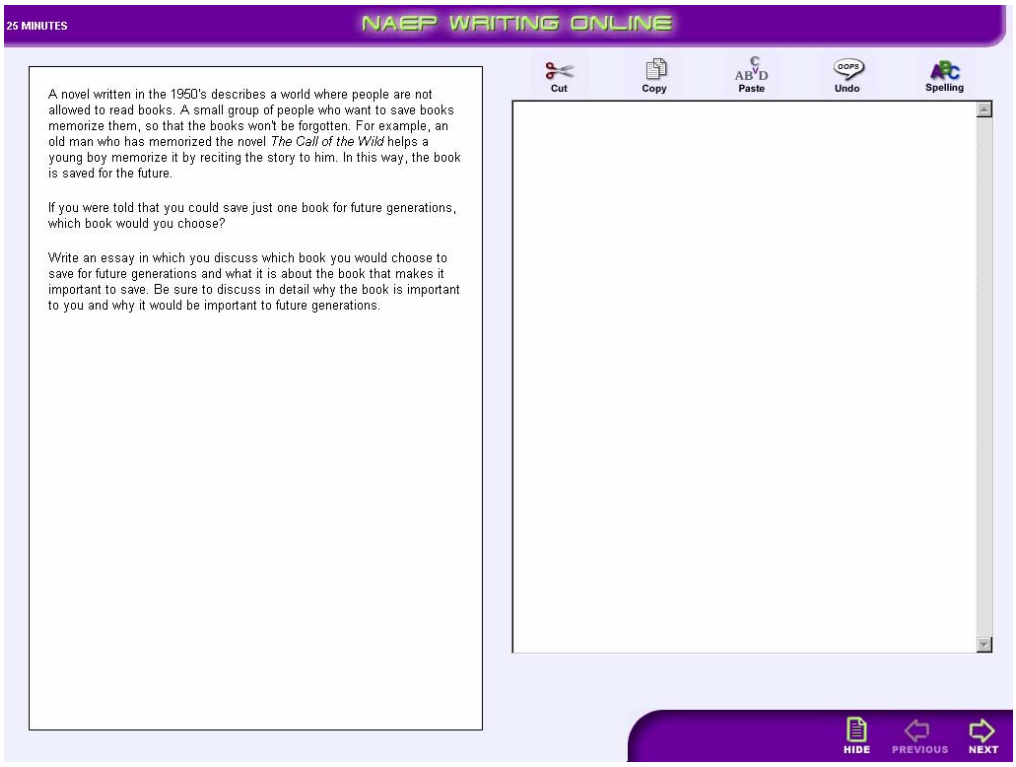
Why should we care about the comparability of scores across delivery modes? We should care because if delivery mode affects scores, our ability to draw valid conclusions from test results may be reduced under some circumstances. For example, it may be reduced:

- if we want to compare results over time and the delivery mode has changed from paper to computer;
- if we want to compare individuals--or aggregate results across them--when some individuals have taken the test on paper and some have taken it on computer (especially if the assignment to delivery mode was not voluntary); or
- if we want to compare individuals (or groups) taking the test on computer and computer delivery affects one type of individual more than another, such that the difference in scores between two individuals (or groups) suddenly becomes larger or smaller than it was for paper testing.

Under any of these circumstances, the conclusions we draw from test results may be wrong. Recognizing this fact, and knowing that NAEP would eventually need to assess writing proficiency on computer, the National Center for Education Statistics commissioned the Writing Online (WOL) study (Horkay, Bennett, Allen, & Kaplan, 2005). This study administered the same two essay questions to one group of students on paper and to another comparable group on computer. Both groups were selected to be nationally representative samples of 8th grade students (i.e., about 13 years old).

The students taking the online writing test used a simplified word processor to enter their responses. This simplified word processor contained functions for cutting, pasting, and copying text; for undoing the most recent action; and for checking spelling (see Figure 2).

Figure 2: Screen Shot of the Writing Online Computer Interface



Note. A writing prompt is on the left and a simplified word processor is on the right. From Horkay et al., (2005).

It's worth noting that, if a student is a good typist who often uses similar word processing tools to write and revise, he or she might write a better essay using this simplified word processor than using pencil and paper. On the other hand, if the student can't type, the essay written here may well be worse than if it had been composed in the traditional way.

To find out how experienced students were with computers, the group taking the online version of this essay-writing test completed three indicators of "computer familiarity." Two of the computer-familiarity indicators were based on students' responses to questions about, in one case, their extent of general computer use and, in the other case, their frequency of using the computer for writing. The third indicator was a hands-on measure that included typing speed (the number of words typed in two minutes from a 78-word passage), typing accuracy (the sum of errors made in typing the passage), and text editing (the number of editing tasks completed correctly, including deleting, inserting, modifying, and moving text).

Finally, the group taking the online writing test also had previously taken a paper writing test. This paper test consisted of two essays that were different from the ones employed in the online test.

What did the nationally representative sample of 8th-grade students taking the online writing test report about the extent to which it used computers for writing? In 2002 when these data were collected, 93 percent of students reported using a computer *at least to some extent* to write. But substantial and almost equal numbers said *to a large extent* (30%) vs. *to a small extent* or *not at all* (29%).

Did 8th-grade students perform differently on the paper and computer writing tests? This study found no significant difference in scores between students taking the writing test on paper and the group taking the same two essays on computer. That result is logically consistent with the distribution of computer use for writing just cited, where a substantial proportion of students used the computer to a large extent and an almost equal proportion used it hardly at all.

Was computer familiarity associated with online test performance? This study found that the hands-on measure of computer familiarity significantly predicted online writing score, after controlling for paper writing performance. The greater the computer familiarity, the higher was the online writing score. To give a sense of the strength of this association, for the same paper writing performance, a student with computer familiarity one standard deviation below the mean would be predicted to get a computer writing score almost one point lower on a 1-6 scale than a student having computer familiarity one standard deviation above the mean. Larger differences in computer familiarity between students with the same paper writing proficiency would be associated with correspondingly bigger discrepancies in computer writing scores.

If We Can Score Student Writing Automatically, Should We Care How the Machine Does it?

Regardless of delivery mode, the need to process results efficiently is important because student writing is time consuming and costly to grade. Assuming that the comparability issues just

described can be dealt with, computer-based testing offers a potentially significant efficiency advantage. Because students taking online writing assessments enter their responses in digital form, those responses can be scored automatically.

Several US testing programs already use automated essay scoring operationally. These programs usually employ it in conjunction with human examiners when the assessment is for making high-stakes decisions. No human examiner is used when the decisions are of less consequence. Testing programs that employ automated essay scoring include the Graduate Management Admission Test (GMAT) Analytical Writing Assessment, ACCUPLACER: WritePlacer Plus and COMPASS e-Write (used by US postsecondary institutions for placement in remedial writing courses), and the Indiana English 11 End-of Course Assessment, which is used for certifying public schools as outstanding and, optionally, as one piece of evidence in assigning course grades to secondary school students.

There are at least four commercially available automated essay scoring systems including PEG (Project Essay Grade) (Measurement, Inc.), IEA (Intelligent Essay Assessor) (Pearson Knowledge Technologies), Intellimetric (Vantage), and e-rater (ETS). In general, the scores generated by these programs agree about as well with human ratings as the human ratings agree among themselves (see Keith, 2003, for a review).

How does a machine score essays? The same general process is used by each of the commercially available programs. First, the developers identify relevant text features that can be extracted by computer (e.g., the similarity of the words used in an essay to the words used in high-scoring essays, the average word length, the frequency of grammatical errors, the number of words in the response). Next, they create a program to extract those features. Third, they combine the extracted features to form a score. And finally, they evaluate the machine scores empirically.

But even if its agreement with human scores is good, does it matter *how* the machine does the scoring? For instance, does it matter how the developers choose to combine the features extracted by the machine to form scores? The method that appears to be used by each of the commercial programs is a brute-empirical one. That is, they use a weighted combination of features that best predicts the scores human judges would assign to those essays under operational grading conditions. Would writing experts choose those same combinatorial weights?

In a study funded by the National Center for Education Statistics, we asked writing experts to weight according to their best judgment 12 features extracted by one automated essay scoring program (Bennett & Ben-Simon, 2005). The essays were taken from the NAEP Writing Online investigation described above. For purposes of the analysis, we grouped the features into five dimensions cited in previous research with the scoring program: Grammar, usage, mechanics, and style; Organization and development; Topical analysis (content); Word complexity; and Essay length. Two committees of experts were asked to weight each dimension independently on a 0-100 scale. These judgments were made by committee members in the abstract, without knowing how the automated essay scoring program worked.

Interestingly, the two committees agreed with one another to a remarkable extent on the relative importance of these dimensions in defining quality for essays written by US 8th grade students. Equally important, the committees *disagreed* to a remarkable extent with the empirical weights assigned by the automated scoring system. The expert committees believed that over 60% of the essay score should be based on the combination of Organization and development and Topical analysis. The empirical weights, in contrast, gave only about *20%* of the emphasis to these dimensions. What received most of the empirical weight? The combination of Grammar, usage, mechanics, and style and Essay length received almost *70%* of it. Yet the committees had awarded only 20% to 26% of the weight to the combination of these dimensions.

As noted, these judgments were made in the abstract before the mechanics of the automated scoring had been described. After both committees received information about the way in which the dimensions were measured, and after one committee saw the empirical weights, the judgment process was repeated. The committee that saw the empirical weights moved closer in its judgments to those weights. However, each committee still gave less emphasis to Grammar, usage, mechanics, and style, and to Essay length, than did the empirical weights. Similarly, each committee gave more consideration to Organization and development and to Topical analysis than did the machine scoring. Finally, scores generated using the "empirically-informed" committee weights proved to be about as meaningful as scores based solely on the empirical weights, suggesting that moderating the optimal weights does not necessarily lead to a reduction in scoring accuracy.

Conclusion

What lessons about technology and writing assessment did we learn from the NAEP studies described in this paper? The first lesson is that scores from paper and computer writing tests may mean different things because computer-based tests may directly measure not just writing skill but, for some students, computer facility. The implication of this finding is that, as long as significant numbers of students write better in one or the other mode, the method by which students are tested may produce different group proficiency estimates. For example, testing students in the mode in which they typically write may lead to different findings than assessing all students in either mode alone.

The second lesson is that how automated scoring is done matters! The weighting of text features derived by an automated scoring system may not be the same as the one that would result from the judgments of writing experts. Statistically optimal approaches may "work" in a predictive sense but they may not work scientifically or educationally, especially if the dimensions students must concentrate on to improve scores are not the ones that writing experts most value.

References

Bennett, R. E., & Ben-Simon, A. (2005). *Toward theoretically meaningful automated essay scoring.* Unpublished final project report, Educational Testing Service.

Horkay, N., Bennett, R. E., Allen, N., & Kaplan, B. (2005). Online assessment in writing. In B. Sandene, N. Horkay, R. E. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje

(Eds.), *Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project* (NCES 2005-457). Washington, DC: National Center for Education Statistics, US Department of Education. Retrieved March 9, 2006 from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457

Keith, T. Z. (2003). Validity of automated essay scoring systems. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

Schleicher, A. (2006). *Are students ready for a technology-rich world? What PISA studies tell us*. Paris, France: Organization for Economic Co-operation and Development (OECD). Retrieved March 6, 2006 from http://www.pisa.oecd.org/dataoecd/28/53/35996337.ppt