Research Memorandum

Moving the Field Forward: Some Thoughts on Validity and Automated Scoring

Randy Elliot Bennett

Research & Development

January 2004
RM-04-01

Moving the Field Forward: Some Thoughts on Validity and Automated Scoring

Randy Elliot Bennett

ETS, Princeton, NJ

January 2004

Abstract

The automated scoring of open-ended items is currently used for licensure tests in medicine and architecture and for tests employed in admissions to graduate management programs. Whereas automated scoring appears to have achieved some of the efficiency goals of its progenitors, it has yet to realize fully its potential to improve the quality of assessment through finer control of underlying constructs. This paper offers thoughts for moving the field forward by, first and foremost, remembering that it's not only the scoring. That is, the hard work is in devising a scoring approach—and more generally, an assessment—grounded in a credible theory of domain proficiency. We can also move the field forward by conducting rigorous scientific research that helps build a strong validity argument; publishing research results in peer-reviewed measurement journals where the quality can be critiqued from a technical perspective; attempting to use competing scoring approaches in combination; employing multiple pieces of evidence to bolster the meaning of automated scores; and, finally, using automated scoring to fundamentally alter the character of large-scale testing in ways that bring lasting educational impact.

Key words: Automated essay scoring, automated scoring, online assessment, computerized testing, computer-based testing

Acknowledgements

I thank Anat Ben-Simon, Isaac Bejar, Henry Braun, Dan Eignor, Bob Mislevy, and David Williamson for their comments on earlier drafts of this manuscript.[1]

What do we need to do to move the field of automated scoring forward? Before answering that question, we might consider what factors motivated automated scoring in the first place. The primary factor has, without question, been efficiency. Scoring by human judges is costly and slow. The advent of Web-based online scoring—that is, where judges read scanned or typed responses via the Internet—appears to have reduced the cost and increased the speed with which test results can be returned to users. There is no costly and time-consuming movement of paper and no gathering, feeding, or housing of judges. Several major testing companies now commonly use Web-based online scoring as a routine component in response processing because they believe it to be an efficiency improvement ("ePEN," 2003; "ETS Online Scoring Network," 2003).

Automated scoring can make similar efficiency claims. For the Graduate Management Admission Test®, for example, the use of automated scoring allows the program to employ one human rater instead of the two it formerly utilized (Enbar, 1999). For other programs, like the National Council of Architectural Registration Boards' Architect Registration Examination, no human judge is employed (except for quality control), so scoring could, in principle, be immediate.

The claim that automated scoring has achieved efficiency improvements can be most reasonably made under two conditions.[2] The first condition is that the process required for preparing the automated system to score responses to new test items is rapid and inexpensive to implement. The second condition is that the examinee volume is large. The greater the examinee volume and the more efficient the scoring preparation for new items, the more competitive that scoring will be vis-à-vis human raters.

Although automated scoring may have achieved, at least in some cases, the efficiency goals of its progenitors, it has yet to achieve fully a potentially more important goal. This goal was, in the beginning, a secondary, almost incidental one. Automated scoring, as Bennett and Bejar (1998) pointed out, allows for relatively fine construct control. That is, the assessment designer can implement through the scoring program the particular set of response features, and a feature weighting, thought to best elucidate the construct of interest. The construct can be tuned by removing features or reweighting them at will. Once the best set of features and weights has been selected, it will be implemented consistently. Human scoring, in contrast, affords only gross control because it is more difficult for humans to deal with multiple features

simultaneously, to weight them appropriately, and to apply scoring rules consistently (Bejar, Mislevy, & Williamson, in press; Dawes, 1979).

As noted, we haven't yet achieved that fine construct control to the same extent as our original efficiency goal. Here are some thoughts about how we might approach the former goal more effectively.

Automated scoring needs to be designed as part of a construct-driven, integrated system. In an article titled, "Validity and Automated Scoring: It's Not Only the Scoring," Bennett and Bejar (1998) argued that a computer-based test was a system consisting of a construct definition, test design, and task design; examinee interface; tutorial; test development tools; automated scoring; and reporting. Further, they noted that the interplay among system components needed to be accounted for in scoring-program design and validation because these components affect one another, sometimes in unanticipated ways. Well, it's still not only the scoring!

One of the key contributions of evidence-centered design (ECD) is that it offers a strong conceptual framework for driving the development of complex assessment systems in a construct-centered way (Mislevy, Almond, & Lukas, 2003). The models emerging from an ECD approach to assessment design—student, evidence, and task—can drive the other components of the assessment also—the user interface, tutorial, test developer tools, reporting formats, tasks, and automated scoring—so that all the components of a computer-based test work synergistically.

In other words, the ECD approach suggests creating scoring as part of the design and validation of an *integrated* system. Such an approach is very different from the one that has often been taken, especially with automated *essay* scoring. Here, automated routines were typically created as general modules to be used with testing programs that already included essay components. In such a situation, the existing test provides the context into which the automated scoring system must fit. The construct definition (such as it may be), the examinee interface, the tasks, the rubric, and the human-generated operational scores become a target for the automated scoring system to hit *regardless* of how well-supported the underlying construct might be.

In honoring Harold Gulliksen, Messick (1989, p. 13) noted that "if you do not know what predictor and criterion scores mean, you do not know much of anything in applied measurement." When we uncritically use human scores as the development model and validation criterion for automated programs, we find ourselves in almost that situation. If we

successfully model the scores of a pair (if we're lucky) of human judges devoting (at most) a few minutes per essay, we also model any systematic biases humans bring to the scoring enterprise. This is an odd result, for automated scoring offers the opportunity to *remove* those biases. Why institutionalize them instead through our modeling methodology?

Further, if we do this modeling through step-wise regression—which some scoring mechanisms have routinely used—we cede control, at least to some degree, of the construct to whatever variables best model the particular raters that happened to score that prompt and the particular examinees that happened to respond to it. We let a statistical rather than a substantive criterion decide which variables to choose and how to weight them in determining examinee performance.

A far more satisfying approach would be to go back to first principles, using ECD methods to define the characteristics of good performance in some relevant context (e.g., secondary school), and then design tasks, rubrics, and automated scoring mechanisms built around those characteristics.[3] In this approach, human scores are no longer the standard to be predicted. Rather, once ECD has been used to define the evidence needed to support claims for proficiency, human scoring is used only to check that the machine scoring is identifying and accumulating that evidence correctly.

I should be clear that the ECD approach does not forsake human judgment. It only uses it differently. That is, human judgment is employed to help define the characteristics of good domain performance (e.g., in writing, architectural design, medical problem solving), the behaviors that would be evidence of it, and the characteristics of tasks that would elicit those behaviors. With respect to scoring, human experts should do what they do best: choose the features of responses to include and their directional relationships to the underlying proficiency (Dawes, 1979). Experts may even be used to determine how to weight each feature because expert weighting may be more easily justified to disciplinary communities (e.g., writing instructors) than empirical weighting and because experts generally produce weights that are nearly as good as empirically derived ones (Dawes, 1979).[4] Thus, in ECD, the emphasis on human judgment is *up front* in design and not solely after the fact as an outcome to be predicted.

In defining the characteristics of good domain performance, we should be sure to take advantage of cognitive research on what makes for proficiency in that domain. The work on automated essay scoring, for example, has capitalized well on theory in computational linguistics

(e.g., Burstein, Kukich, Wolff, Lu, & Chodorow, 1998) and on the psychology of knowledge representation (e.g., Landauer & Dumais, 1997). But there has been little, if any, cognitive *writing* theory incorporated in any of the automated essay scoring work even though these systems typically issue scores for writing proficiency.

Automated-scoring research needs to use appropriate methods. Although some applications of automated scoring have incorporated measurement methods effectively, in other applications the research has not been as rigorously done. Rigor is particularly important for automated scoring because without scientific credibility, the chances for general use in operational testing programs are significantly diminished, especially with the recent emphasis on scientifically based research as a prerequisite for the purchase and use of educational programs and products (Feuer, Towne, & Shavelson, 2002). Such rigor is essential whether the research is done in an experimental setting or is conducted in the context of an operational testing program. The methodological weaknesses include the most basic rudiments of scientific investigation. Some studies fail to describe the examinee population, the sampling method, the evaluation design, how the automated scoring program works, the data analysis methods, or even the results adequately enough to assess the credibility of the work. Other flaws are, perhaps, more subtle. These include reporting results from the sample used to train the scoring program rather than from a cross-validation; using only a single prompt, which offers little opportunity for generalization to *any* universe of tasks; failing to correct for chance agreement, which can be very high when scoring scales are short and the effective range used by human raters is shorter still; combining results across grade levels, which produces spuriously high correlations; reporting findings only for a general population without considering subpopulations; and reporting only exact-plus-adjacent agreement because this is the standard traditionally used for human rating.

Perhaps the most subtle but pernicious flaw is mistaking machine-human agreement for validation. Machine-human agreement is but a piece of evidence in the validity argument (Bennett & Bejar, 1998; Williamson, Bejar, & Hone, 1999). For automated scores, as for any other scores, the validity argument must rest on an integrated base of logic and data, where the data allow a comprehensive analysis of how effectively those scores represent the construct of interest and how resistant they are to sources of irrelevant variance (Messick, 1989). Exemplary in this regard, at least in automated essay scoring, is the work of Powers and colleagues, which

represents the only comprehensive comparison of human vs. automated scoring vis-à-vis *multiple* external criteria (Powers, Burstein, Chodorow, Fowles, & Kukich, 2001), as well as the only analysis of the extent to which automated graders can be manipulated into giving scores that are either undeservedly high or low (Powers, Burstein, Chodorow, Fowles, & Kukich, 2000). Studies like these set a standard toward which automated scoring research should strive because they illustrate that it is *validity*—and not just rater agreement or reliability more generally—that should be the standard against which alternative scoring systems should be compared.

Automated-scoring research needs to be published in peer-reviewed measurement journals. The results of automated scoring research in medical licensure, mathematics, computer science, and architectural design have been widely published in peer reviewed measurement journals (e.g., Bejar, 1991; Bennett et al., 1990; Bennett & Sebrechts, 1996; Bennett, Sebrechts, & Rock, 1991; Clauser et al., 1995; Clauser, Margolis, Clyman, & Ross, 1997; Williamson et al., 1999). Perhaps because of commercial pressures or differences in the disciplinary communities responsible for development, the same cannot be said for automated essay scoring.[5] In this field, the research has appeared as conference papers, in conference proceedings, in linguistics journals, as book chapters, and as technical reports. These dissemination vehicles, while certainly respectable, don't typically get the same level of methodological scrutiny as the peer-reviewed measurement journals. That may be one reason why the automated essay scoring research, on average, hasn't been as rigorous from a measurement perspective as that in other fields. A fair amount of the automated essay scoring research is quite good and would only be improved through such publication.

Future research should study the use of multiple scoring systems. The basic idea behind using multiple sources of evidence—be they items, raters, or occasions—is that systematic, valid sources of variance should cumulate and unsystematic, irrelevant ones should wash out (Messick, 1989). Thus, we often have multiple human raters grade the same student productions with the knowledge that, over the long run, the resulting scores will be more generalizable than if they had been rendered by a single judge. We can use the same principle with scoring programs in the hope that different approaches would complement one another by balancing weaknesses—that is, systematic but irrelevant and idiosyncratic sources of variance. We should not expect this approach to work if the automated graders are simply trivial variations of one another. In that case, they will share similar strengths and weaknesses. The greatest chance for success would

seem to come from systems that are substantially different in their scoring approaches. In a successful multiple-system approach, the human grader might be able to play a quality-control role, checking only a sample of the machine-scored responses, as is done in the Architect Registration Examination, rather than rating all responses as a human still does for the essay section of the Graduate Management Admission Test.

Automated scoring could help change the format and content of assessment. Large-scale assessment currently depends very heavily on one-time, multiple-choice tests. The reasons we use multiple-choice tests so widely are because they can be efficiently scored and because they produce a great deal of generalizable information in a short period. However, educators typically disdain them because they poorly reflect the format, and often the content, students must learn to succeed in a domain. That is, educators see multiple-choice questions as too distant in format and substance from the activities routinely undertaken in the study of school achievement domains. Because they are not closely tied to curriculum, such tests tend to offer little information of immediate use to students and teachers. As a result, we employ them largely for external accountability purposes and administer them as infrequently as possible to limit the disruption of classroom learning.

In conjunction with the Internet, automated scoring has the potential to lighten the influence of the one-time, multiple-choice testing event and better integrate assessment into the curriculum. The Internet provides the means to easily deliver standardized performance assessment tasks to the classroom (Bennett, 2001). Those tasks can be as simple as an essay prompt or they can be more complex, incorporating such dynamic stimuli as audio, video, and animation. Tasks can also be highly interactive, changing how they unfold depending upon the actions that the student takes—much as the simulation section of the United States Medical Licensing Examination does today (United States Medical Licensing Examination, 2003). They could also be designed so that they act—perhaps incidentally, perhaps explicitly—as learning experiences. Such tasks should be very attractive to educators and policy makers, as long as student performance can be automatically scored.

One could imagine such tasks being used in several ways. One way is simply as a means of allowing performance tasks to be used more often in one-time testing events such as traditional standardized tests. Performance tasks, like essays, are employed now to a limited degree at great expense and with substantial delay in reporting. Automated scoring has the

potential to reduce the cost and delay and thereby increase the attractiveness of using performance tasks as a component in one-time testing systems.

A second way is formatively; that is, as frequent practice, progress monitoring, or skill diagnosis measures, all in preparation for some culminating one-time assessment. A third possibility is to add a summative layer to this formative use by taking information from each periodic progress measure to supplement the data provided by the culminating assessment. A last possibility is to, at some point, do away with the culminating assessment entirely and let the periodic events bear the full formative and summative weight (Bennett, 1998).

Whether or not it is used in combination with a culminating assessment, any summative use of periodic performance assessment would require tasks built to measure critical domain proficiencies, enough tasks to provide generalizable results, and an evidence model capable of aggregating information across tasks in a meaningful way. Creating such assessments would not be cheaper than creating multiple-choice tests, although the intellectual discipline required for automated scoring should help make task creation more systematic and efficient (Mislevy, Steinberg, & Almond, 2002). In any event, the educational benefits might well outweigh the costs.

Conclusion

To sum up this commentary, we can move the field forward in several ways. First, we can remember that it's *still* not only the scoring. The hard work in automated scoring is not in mimicking the scores of human raters; it's in devising a scoring approach—and more generally, an assessment—grounded in a credible theory of domain proficiency, be that domain writing, mathematics, medical problem solving, or architectural design. Second, we can conduct rigorous scientific research that helps build a strong argument for the validity of the scores our automated routines produce. Third, we can publish the results of our research in peer-reviewed measurement journals where the quality of our work is critiqued from a measurement perspective and, hopefully, improved as a result. Regardless of the techniques we use—natural language processing, neural networks, statistical methods—automated scoring is first and last about providing valid and credible *measurement*. Without incorporating measurement principles and submitting our methods and results to technical review, automated scoring is not likely to be either valid or credible. Fourth, we can attempt to use competing scoring approaches in combination, employing multiple pieces of evidence to bolster the meaning of automated scores.

Finally, we may be able to use automated scoring to create assessments that are more closely tied to the format and content of curriculum. Perhaps through this use of automated scoring, we can fundamentally alter the character of large-scale testing in ways that bring lasting educational impact.

References

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76*, 522-532.

Bejar, I. I., Mislevy, R. J., & Williamson, D. M. (in press). Human scoring. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing*.

Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing* (Policy Information Center Perspective; ETS RR-97-14). Princeton, NJ: ETS, Policy Information Center. Retrieved September 15, 2003, from ftp://ftp.ets.org/pub/res/reinvent.pdf

Bennett, R. E. (2001). How the Internet will help large-scale assessment reinvent itself. *Education Policy Analysis Archives*, *9*(5). Retrieved September 15, 2003, from http://epaa.asu.edu/epaa/v9n5.html

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9-17.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C. & Soloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement, 14*, 151-162.

Bennett, R. E., & Sebrechts, M. M. (1996). The accuracy of expert-system diagnoses of mathematical problem solutions. *Applied Measurement in Education, 9,* 133-150.

Bennett, R. E., Sebrechts, M. M., & Rock, D. A. (1991). Expert-system scores for complex constructed-response quantitative items: A study of convergent validity. *Applied Psychological Measurement, 15*, 227-239.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). *Enriching automated essay scoring using discourse marking and discourse relations*. Paper presented at the 17th International Conference on Computational Linguistics, Montreal, Canada.

Clauser, B. E., Margolis, M. J., Clyman, S. G., & Ross, L. P. (1997). Development of automated scoring algorithms for complex performance assessments: A comparison of two approaches. *Journal of Educational Measurement, 34,* 141-161.

Clauser, B. E., Subhiyah, R. G., Nungenster, R. J., Ripkey, D. R., Clyman, S. G., & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgment process of experts. *Journal of Educational Measurement, 32,* 397-415.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571-582.

Enbar, N. (1999). This is e-rater. It'll be scoring your essay today. *Business Week Online*. Retrieved September 15, 2003, from http://www.businessweek.com/bwdaily/dnflash/jan1999/nf90121d.htm

ePEN electronic performance evaluation network. (2003). Retrieved September 15, 2003, from the Pearson Educational Measurement Web site: http://www.pearsonedmeasurement.com/epen/index.htm

ETS online scoring network. (2003). Retrieved September 15, 2003, from http://www.ets.org/reader/osn/

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher, 31*(8)*,* 4-14.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (third edition). New York: MacMillan.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (ETS RR-03-16). Princeton, NJ: ETS.

Page, E. B. (1967). The imminence of grading essays by computer. *Phi Delta Kappan, 48,* 238-243.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2000). *Comparing the validity of automated and human essay scoring* (ETS RR-00-10). Princeton, NJ: ETS.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). *Stumping e-rater™: Challenging the validity of automated essay scoring* (ETS RR-01-03). Princeton, NJ: ETS.

United States Medical Licensing Examination. (2003). Preparing for the test. *2004 USLME Bulletin of Information.* Retrieved September 15, 2003, from http://www.usmle.org/bulletin/2004/preparingforthetest.htm

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). 'Mental model' comparison of automated and human scoring. *Journal of Educational Measurement, 36*, 158-184.

Williamson, D. M., Mislevy, R. J., &Bejar, I. I. (Eds.). (in press). *Automated scoring of complex tasks in computer-based testing.* Hillsdale, NJ: Erlbaum.

Notes

[1] This Research Memorandum is based on a chapter that will appear in Williamson, Mislevy, & Bejar (in press).

[2] This analysis presumes that the test is already electronically delivered and that the question of interest revolves around the incremental cost of automated vs. traditional scoring and *not* around the incremental cost of computer vs. paper delivery.

[3] Iterative cycles would be anticipated, of course, in a continuing interplay among rational analysis, conjectures about construct-based elements of an assessment system, and empirical tests and model criticism (R. Mislevy, personal communication, September 10, 2003). This is the basis of any good validation program.

[4] A second problem with weights derived through regression is that the best linear composite invariably puts greater weight on those features that work most effectively in prediction for the majority of candidates. It is possible that some well-written essays will emphasize features differently than a brute empirical weighting says they should and, thus, get lower machine scores than an astute human judge might give (H. Braun, personal communication, October 10, 2003).

[5] Ironically, work on automated essay scoring predates all other forms of automated scoring by several decades (e.g., Page, 1967).