Automated Scoring of Constructed-Response Literacy and Mathematics Items

RANDY ELLIOT BENNETT

JANUARY 2011

Executive Summary

The Race to the Top assessment consortia have indicated an interest in using "automated scoring" to more efficiently grade student answers to constructed-response literacy and mathematics tasks. Automated scoring refers to a large collection of grading approaches that differ dramatically depending upon the constructed-response task being posed and the expected answer. Even within a content domain, automated scoring approaches may differ significantly, such that a single characterization of the "state of the art" for that domain would be misleading.

This white paper identifies potential uses and challenges around automated scoring to help the consortia make better-informed planning and implementation decisions. If the benefits sought from constructed-response tasks are to be realized, automated scoring must be implemented and used thoughtfully; otherwise, automated scoring may have negative effects strikingly similar to those commonly attributed to multiple-choice tests.

This paper concentrates on tasks that call for complex constructed responses not amenable to scoring via exact-matching approaches. In literacy, three task types and approaches to scoring are described: (1) the essay, (2) the short text response, and (3) the spoken response. Of these three, automated scoring for the essay and for a subclass of spoken response has most often been used in high-stakes testing programs.

For mathematics, four task categories are examined: (1) those calling for equations or expressions, where the problem has a single mathematically correct answer that can take many different surface forms; (2) those calling for one or more instances

Dr. Randy Elliot Bennett[1] is Norman O. Frederiksen Chair in Assessment Innovation at Educational Testing Service (ETS), a nonprofit organization that conducts educational research and operates testing programs for use in education and work environments.

This paper was produced in partnership with Arabella Advisors. Arabella Advisors is a philanthropy consulting firm supporting the efforts of individual, family, institutional, and corporate donors worldwide. We are committed to unbiased analysis that helps donors support issues and nonprofits with confidence. Our expertise and insights transform philanthropic goals into results. For more information on our firm, please visit our website at www.arabellaadvisors.com.

from a potentially open-ended set of numeric, symbolic, or graphical responses that meet a given set of conditions; (3) those requiring the student to show the symbolic work leading to the final answer; and (4) those asking for a short text explanation. For various reasons, none of these task classes has been regularly used in consequential assessments.

The use of automated scoring in high-stakes testing programs poses important issues, some of which transcend content domain. One issue is that the components of a computer-based testing system affect one another such that a poorly designed student interface, for example, may reduce the accuracy of automated scoring. A second issue is that some automated scoring approaches are developed to predict the grades produced by operational human raters (and are evaluated primarily on the accuracy of that prediction), even though it is not generally known how faithfully operational raters actually follow scoring rules. Finally, the increased complexity of automated methods, and nondisclosure of technical details by some vendors, means that users cannot always know, or defend, how the scoring works.

To ensure that automated scoring helps the Race to the Top assessment consortia achieve their innovation goals, seven recommendations are offered below.

RECOMMENDATIONS

1. Design a computer-based assessment as an integrated system in which automated scoring is one in a series of interrelated parts
2. Encourage vendors to base the development of automated scoring approaches on construct understanding
3. Strengthen operational human scoring
4. Where automated systems are modeled on human scoring, or where agreement with human scores is used as a primary validity criterion, fund studies to better understand the bases upon which humans assign scores
5. Stipulate as a contractual requirement the disclosure by vendors of those automated scoring approaches being considered for operational use
6. Require a broad base of validity evidence similar to that needed to evaluate score meaning for any assessment
7. Unless the validity evidence assembled in Number 6 justifies the sole use of automated scoring, keep well-supervised human raters in the loop

Automated Scoring of Constructed-Response Literacy and Mathematics Items

Automated scoring refers to a large collection of grading approaches that differ dramatically depending upon the constructed-response task being posed and the expected answer. Even within a content domain, automated scoring approaches may differ significantly, such that a single characterization of the “state of the art” for that domain would be misleading.

There are at least two general classes of assessment task for which automated scoring can be used. The first class entails constructed-response tasks that can be graded using exact-matching techniques. For these problems, the scoring challenge is relatively trivial: the correct answer(s) are known in advance and can be used to evaluate the quality of the student’s response.

A second general class consists of problems for which the responses are too complex to be graded in the aforementioned way. Automated scoring of complex responses is generally accomplished via a scoring “model.” The model extracts features from the student response and uses those features to generate a score. Tasks from this second class may be scored as right or wrong, but in many cases they can be graded on a partial-credit scale. When assessment professionals think of “automated scoring,” they are usually referring to tasks from this class. As a result, the remainder of this paper focuses on this task type.

The structure of the paper is as follows: The first section describes what automated scoring is, the types of tasks with which it is currently used, and how such scoring generally works. Next, the issues that are posed by the use of this technology in high-stakes assessment are discussed. The paper closes with detailed recommendations for how the consortia might use automated scoring responsibly and effectively.

Automated Scoring of Complex Constructed Responses

Literacy

In literacy, three task types warrant consideration: the essay, the short text response, and the spoken response. For both essay and short text responses, students must enter their answers via computer. (Recognition of handwritten responses has not yet reached the point where use in a consequential testing program is defensible.)

In automated essay grading, the machine uses a set of low- to mid-level computable text characteristics to predict the score that a human rater would be likely to assign. This approach has been quite successful; across many studies and different automated grading programs, scores computed in this way have typically been found to agree as highly with human scores as two human raters agree with one another.[2] Automated essay scoring is used in several high-stakes testing programs, usually in conjunction with a human rater.

It is important to note that, regardless of what vendors may claim, a machine does *not* read, understand, or grade a student's essay in the same way as we like to believe that a human rater would. It simply *predicts* the human rater's score. This fact has two critically important implications. The first implication is that concentrating instruction on the features measured by automated essay scoring is likely to improve a student's score as well as certain lower-level writing competencies. But such concentration is unlikely to improve student writing with respect to higher-order competencies such as the quality of ideas, argumentation, audience awareness, and rhetorical style, which no extant automated scoring approach directly measures. The second implication is that, despite the inclusion of various fail-safe mechanisms, when used alone automated essay scoring may be subject to manipulation.[3]

It is important to note that, regardless of what vendors may claim, a machine does *not* read, understand, or grade a student's essay in the same way as we like to believe that a human rater would. It simply *predicts* the human rater's score.

Short text responses for literacy assessment are typically associated with questions calling for factual information, analysis, and explanation or conceptual understanding, among other things. For some types of short text responses, automated scores agree with human scores as often as human scores agree with one another.[4] For other types of short text responses, automated scoring performs quite poorly relative to the performance of human raters. The degree to which automated scoring works effectively depends almost entirely on how predictable and limited in semantic, lexical, and syntactic variety student responses turn out to be. The more limited the response variety, the more likely automated scoring is to be effective. Limiting response variety depends, in turn, on crafting the test questions to enforce that restriction. Enforcing that restriction too severely, of course, narrows the type of problem solving that can be assessed, and narrowing problem solving may run counter to the purpose of using constructed-response questions in the first place. As a result, aside from tasks that can be scored by exact match, automated scoring of short text responses has not generally been used in high-stakes literacy assessment programs.

The third class of literacy task entails spoken responses. These responses are "entered" by speaking into a computer via a microphone or telephone. Speaking tasks are typically segmented into low-entropy and high-entropy tasks. Low-entropy tasks are ones for which the responses are fairly predictable. Examples include oral reading from a printed passage, repeating an orally presented stimulus, giving an answer to a highly constrained factual question (e.g., "What season comes just after winter?"), or describing a simple picture (e.g., a child throwing a ball). High-entropy tasks, in contrast, produce unrestricted, spontaneous speech.

The general approach used in automated speech scoring is similar to that employed in automated essay scoring. A significant difference is that, in speech scoring, the response takes the form of a digitized sound file with unknown word identities. Therefore, a machine recognizer must first process the response and generate hypotheses for each word. Because of insufficient recognizer accuracy (especially for non-native speakers), as well as semantic challenges similar to those generally encountered in extracting meaning from text, automated scoring is rarely used in consequential testing programs. When it is employed, its use has been with low- to medium-entropy, rather than high-entropy, speech tasks.

Table 1 summarizes the literacy task types. Note that different types can be assembled to form a set, which, tied together by an introductory scenario and culminating purpose, may function much like a performance task.

TABLE 1: A SUMMARY OF AUTOMATICALLY SCORED TASKS IN LITERACY

Task Type	Example Task	Intended Target Competency for the Example Task	Scoring Approach	Scoring Quality	Key Limitations
Essay	*Many students have jobs after school. They may work in a store, make deliveries, or do manual labor. Explain why you feel that working after school is a good or a bad idea. Make sure to give reasons and examples to support your position. You will have 25 minutes to write your response.*	Persuasive writing skill, including argumentation	From an automatic extraction and evaluation of low- to mid-level response features, predict the score that a human rater would be likely to assign.	Across many studies and multiple automated scoring programs, scores of human raters are predicted about as well as one human rater's score predicts another human rater's score.	Measuring only low- to mid-level response features may encourage teachers to focus instruction—and students to center their learning practice—on those features to the exclusion of higher-order writing and critical thinking skills (e.g., argumentation).
Short text response	*Why did Brutus help to kill Julius Caesar? State your answer in two sentences or less.*	Literary knowledge. Skill in clearly expressing that knowledge.	Compare the response to one or more test-developer keys to determine if the response is a paraphrase of the key or can be derived from it by implication.	From the limited data reported in testing contexts, performance is comparable to human scoring when student responses are narrow in semantic, lexical, and syntactic variety, and when the human scoring rubric is clear. Lower quality relative to human scoring when these conditions do not hold. This result is consistent with data from non-testing contexts.	Test questions can be crafted to limit response variety but restricting variety too severely may reduce authenticity, with potentially negative wash-back effects. Misspellings may pose a bigger problem for automated than for human scoring as the scoring program may be unable to infer the intended word.
Low-entropy spoken response	*Read the following sentences out loud:* *1. Trash is a big problem in the city.* *2. Garbage seems to pile up faster than the trucks can pick it up.* *3. The city was going to hire more sanitation workers.* *4. The budget crisis undermined that solution.*	Pronunciation; reading fluency.	Automatically identify the words that compose the response. Extract and evaluate low- to medium-level response features, and use that evaluation to predict the score that a human rater would be likely to assign.	From the limited data reported in testing contexts, scores of human raters appear to be predicted about as well as one human rater's score predicts another human rater's score. This result is consistent with data from non-testing contexts.	Using a speech recognizer trained on student samples different from the population being tested may lower accuracy.
High-entropy spoken response	*Talk about a person who had a positive influence on you. Explain how this person affected you. You will have 15 seconds to prepare your response and 45 seconds to answer.*	Fluency and clarity; vocabulary precision; accuracy of grammar; coherence; accuracy of content.	Automatically identify the words that compose the response. Extract and evaluate low- to medium-level response features, and use that evaluation to predict the score that a human rater would be likely to assign.	From the limited data reported in testing contexts, scores of human raters appear to be predicted considerably less accurately than one human rater's score predicts another human score. This result is consistent with data from non-testing contexts.	Automatic recognizers do not yet have sufficiently high accuracy rates for unrestricted speech. The evaluation of response content is constrained by the same unresolved challenges that limit the scoring quality of short text responses and of essays.

Note: The examples above are generally targeted at the secondary school level.

Mathematics

In mathematics, tasks requiring complex responses fall into at least four categories: (1) those calling for equations or expressions, where the problem has a single mathematically correct answer that can take many different surface forms; (2) those calling for one or more instances from a potentially open-ended set that meets given conditions; (3) those requiring the student to show the symbolic work leading to the final answer; and (4) those asking for a short text response. The accuracy with which the responses to such tasks can be scored varies dramatically. Also highly variable across categories is the extent to which the computer affords an easy and familiar way for students to respond. Finally, the effort needed to create scoring models differs significantly across these four task classes.

For questions requiring the entry of an equation or expression, the problem for the scoring mechanism is one of mathematical paraphrase. That is, given a test developer key (which takes the form of a correct expression or equation), is the student's response mathematically equivalent?[5] Although the scoring of such responses is well within the state of the art, the computer keyboard is not well suited to symbolic entry. Further, few primary, or even secondary, students are familiar with the software interfaces that have been created for such entry and editing (e.g., MathType).[6] If tablet or touch-screen computers such as the iPad take hold in school settings, this situation may change. But, as of this writing, automated scoring for tasks of this type has rarely been used in high-stakes testing programs.

The second task class involves problems calling for one or more instances from a potentially open-ended set that meets given conditions. Depending on the problem, the answer may be posed in numeric, symbolic, figural, or graphical form. These problems are attractive because they don't specify all the information needed to generate an answer. Instead, they require the student to deal with a modicum of ambiguity, something more characteristic of real-world problems than of multiple-choice tests. As a result, for problems in this class there is no single correct answer; instead, there are many good (quantitatively different) answers. A simplistic example would be, "Give a number that is divisible by both 12 and 8." There is an infinite set of such numbers, and the adept student will know how to quickly find one. Infinite set aside, all the automated scoring mechanism needs to do is divide the student's response by 12 and by 8; if the result of both divisions is an integer, the response is correct.

Responses to tasks from this general class would appear to be automatically scorable with high accuracy. Also, aside from symbolic answers, responses can be entered relatively easily. However, such tasks have not been widely studied, nor have they been used with any frequency in consequential testing programs.[7]

The third task class involves questions that ask the student to show the work leading to some final answer—that is, the student's solution process. Responses to such items will usually offer a sequence of steps within one or more correct approaches. Identifying the correct approaches, steps, and sequences, and then programming that analysis into a computer, is a highly complex and labor-intensive process. Because a machine does not have the same degree of background knowledge or level of inferential capability as a human grader, this process needs to be done at a level of detail that far exceeds that required for human scoring of the same responses. As a consequence, the automated scoring of such tasks has not found its way into high-stakes assessments.[8, 9]

The last class calls for a short text response (that can't be scored through an exact match). Such items might be used to assess knowledge of factual information, conceptual understanding, or procedural skill. The item may ask for a text entry alone or, alternatively, as the justification for a companion answer (which itself could be a selected response or a numeric, symbolic, or graphical response). Questions like these may be scored using the same approach as employed for short-text literacy responses. For some short-text mathematics questions, automated scores agree with human scores as often as two human scores agree with one another. For other short-text math questions, automated scoring functions much less acceptably. In general, mathematics questions calling for a text response have not been used in high-stakes assessment programs.

TABLE 2. A SUMMARY OF AUTOMATICALLY SCORED TASKS IN MATHEMATICS.

Task Type	Example Task	Intended Target Competency for the Example Task	Scoring Approach	Scoring Quality	Key Limitations
Equation or expression that can take many different surface structures	*If 12 eggs cost x cents and 20 potatoes cost y cents, what is the cost in cents of 2 eggs and 3 potatoes? Enter an expression in the box below.*	Mathematical modeling of a verbally presented problem situation.	Compare the response to the test-developer key to see if the two entities are mathematically equivalent.	From the data reported and widespread commercial use of the underlying symbol manipulation technology, appears likely to be close in accuracy to multiple-choice machine scoring.	Computer keyboard not optimally suited to symbolic entry and editing; few primary or secondary students are familiar with the specialized interfaces that allow for that entry and editing. May change if touch-screen or tablet computers are widely adopted in school settings.
Instance from an open set in numeric, symbolic, or graphical form	*Jane is riding her bike at a steady rate for one minute, at which point she sees a child up ahead and applies the brakes with constant force, coming to a full stop. Create a plot on the labeled grid below that shows what Jane did.*	Graphical modeling of a verbally presented problem situation.	Compare the response to the stated conditions (i.e., constant non-zero speed for one minute, followed by a linearly decreasing speed to zero miles per hour).	From the limited data reported in post-secondary samples, appears close in accuracy to human scoring.	Each item requires a customized scoring model.
Symbolic work leading to a final answer	*A plumber charges $75 for the first hour of service and $50 for each additional hour. A bill of $225 represents how many hours of the plumber's service? Show the work leading to your answer.*	Problem representation, solution planning, and procedural execution.	Compare the response to one or more sets of pre-defined correct (and perhaps incorrect) solution steps.	Comparable to human scoring in assessment research studies and in widely used commercial intelligent tutoring programs.	Requires iterative, highly labor-intensive development, including cognitive analysis of the solution processes involved in solving each class of items, encoding of that knowledge in a computer program, evaluation of student responses, comparison of scores to those of human judges, and revision of the scoring program.
Short text response	*Most students in Jose's class are tall, and a few are short. Click on the statistic that would best represent the average height of the students in his class.* *o Median* *o Mean* *o Mode* *Explain your choice.*	Conceptual understanding.	Compare the response to one or more test-developer keys to determine if the response is a paraphrase of the key or can be derived from it by implication	From the limited data reported in testing contexts, performance is comparable to human scoring when student responses are narrow in semantic, lexical, and syntactic variety, and when the human scoring rubric is clear. Lower quality relative to human scoring when these conditions do not hold. This result is consistent with data from non-testing contexts.	Test questions can be crafted to limit response variety but restricting variety too severely may reduce authenticity, with potentially negative wash-back effects. Misspellings may pose a bigger problem for automated than for human scoring as scoring program may be unable to infer the intended word.

Note: The examples above are generally targeted at the secondary school level.

Issues

In this section, we discuss three major issues that are posed by the use of automated scoring in high-stakes assessment programs. Some of these issues were briefly introduced above. Although they are each very challenging, there are ways to address them, which are presented as recommendations in the paper's closing section.

It's Not Only the Scoring

In a computer-based performance test, scoring is one component in a highly complex system. The components of that system include: (1) the construct definition, the test design, and the task design, all of which should be closely related to one another; (2) the interface used by the student to take the test; (3) the tutorial for familiarizing students with the interface and the task types; (4) the tools that test developers use to create the items; (5) the automated scoring program(s); and (6) the mechanisms for reporting results. Because these system components affect one another, their interplay must be accounted for in the design and validation of the scoring program. That is, automated scoring can't be separated conceptually—and shouldn't be separated practically—from the other test components.[10]

The dependencies among components are particularly salient with respect to the connection between the computer interface and automated scoring. For a multiple-choice test, all the student needs to know is how to navigate from one question to the next, how to use a limited number of tools (e.g., review screens, an onscreen calculator), and how to select a response. For a computer-based performance test, the situation is far more complicated because the scoring program must be able to analyze whatever the interface allows the student to submit (or, if not, know enough to route the response to a human rater for resolution). The more the interface constrains what the student is able to enter, the easier that response will be to automatically score. But the more the interface constrains the student, the further educators get from measuring the naturalistic problem solving that motivated the use of constructed-response and other performance tasks.

A brief example with respect to short text responses might help clarify the issue. If a math question asks why the median is better than the mean for representing a long-tailed distribution, any response that notes the median's reduced sensitivity to extreme values might be accepted as correct. Equal credit probably would be given to responses regardless of the spelling of "sensitivity;" "senstivity," "sentiviity," or "sensativity" would each do, as long as the meaning of the response in which this misspelling appeared was otherwise clear and correct. In this question, then, spelling knowledge might be considered an irrelevant factor.

For the scoring program, however, spelling errors can be problematic because they introduce pseudo words that the program might not be able to associate with the intended real word. To resolve this situation, one approach might be to try to prevent such errors from ever reaching the scoring program. Errors can be prevented by automatically checking the student's response when it is submitted through the interface and displaying words that are not recognized for the student's review and correction. But exactly how this interaction between spell-checker and student is structured can have a significant effect on test validity. Consider, for example, what might happen to an English language learner if the interface was structured to offer a list of candidate real words for each misspelling it detected. If the student consistently selects the wrong real word, the interface will pass on a set of responses that the scoring program won't be able to match to the key and will grade as incorrect. The automated scoring will have been accurate in the sense that it correctly graded what the interface transmitted. But the test will have failed to fairly and validly evaluate this student's understanding because "it's not only the scoring."

Basing Design on the Operational Scores of Human Raters

Developers of some types of automated scoring have chosen to design their approaches to predict the holistic grades produced by human raters under operational testing conditions. Both automated essay scoring and high-entropy speech scoring are examples. In high-stakes testing programs, human scoring proceeds generally like this: First, raters are trained to implement a scoring rubric that identifies the features to be used in grading student responses; as part of this training, raters practice with sample responses. Next, the raters are qualified by asking them to grade a new set of pre-scored responses, allowing raters to grade operationally only if their performance on the qualification set exceeds some standard. Once they start scoring operationally, their performance is monitored by, among other things, computing their agreement levels with a second operational rater and/or with pre-scored responses seeded into their response queues. If a rater either consistently shows low levels of agreement or grades too slowly, that rater is dismissed.

This system appears to work well in that, for large-scale testing programs, the usual outcome is reasonably high levels of agreement and fast scoring.[11] It is not known, however, how faithfully raters follow the intended rubric under these operational conditions.[12] That is, there is no significant research base that would allow assessment staff to understand exactly what features humans typically use when they score or how they weight those features to arrive at a summary judgment. It is possible, for example, that in operational essay scoring, raters use response features that are quickly discernible (for speed), highly objective (for agreement), and correlated with student performance (for producing a result that approximates what following the rubric would accomplish). Essay length and low-level grammar, usage, mechanics, and style errors might be such features.[13] To the extent it occurs, using such shortcuts (instead of a careful reading of the response) could produce pernicious effects. If students and teachers discover this "true" rubric, they may begin to concentrate their efforts on those same features, diverting attention from the higher-level writing skills they should be practicing (and which operational raters might be ignoring).[14] Just as worrisome is that because developers have chosen to design their automated approaches to emulate the results of operational raters, whatever shortcuts human raters may be using, automated approaches may be unintentionally institutionalizing (e.g., over-weighting low-level essay features).[15] And, of course, it then becomes dangerously circular to use those same operational raters as the primary validation criterion, a practice that is all too common in the field.

Access to How Automated Scoring Works

Because constructed-response questions are highly valued by many constituencies in educational as well as in occupational testing, and because human scoring is so costly, significant resources are being invested in computerizing the scoring process. As a result, technologies for scoring are rapidly improving. That rapid improvement is both a benefit and a liability. The liability is that, as these technologies become more capable, they become harder for test users to understand substantively. They are harder to understand because they are ever more complex and because they increasingly depend on areas of expertise well outside of psychometrics (e.g., natural language processing). Further, to protect their investments, some vendors have chosen not to disclose the details of their methods. The end result of increased complexity and nondisclosure is that neither the state education department staff nor that department's technical advisory committee will be able to independently review or recount *how* scores are generated.

Of course, if it is not known how the automated scoring works, it is not possible to fully judge:

- The fidelity of the resulting scores to the intended construct
- The testing purposes, contexts, and populations for which automated scoring might *not* work and for which sufficient empirical data should be gathered
- How the automated scoring program might be gamed
- How the scoring might negatively impact learning and instruction by affecting the behavior of teachers and students

If judgments like these cannot be made knowledgeably, how can public support for the use of automated scoring for high-stakes testing purposes be effectively garnered or that use be legally defended?

Recommendations

Used responsibly, automated scoring can, in fact, be part of the innovation required by the Race to the Top assessment consortia. To help in that responsible use, here are seven recommendations:

1. ***Design a computer-based assessment as an integrated system in which automated scoring is one in a series of interrelated parts.***

   a. Focus particular attention on the interface design and its interaction with the automated scoring program. Make sure that what the interface allows the student to enter, the automated program can score (or can route to a human rater). Find an appropriate balance between answer-entry constraints and the higher-level problem solving intended by the use of constructed-response tasks.

   b. Conduct usability studies to ensure that students can quickly learn and easily employ the interface to enter their constructed responses. Ensure that students have ample opportunity to learn the interface through using the test tutorial, through practice-test opportunities, and through the use of the same interface in online formative assessment or instructional activities.

   c. Conduct cognitive labs to ensure that the strategies, processes, knowledge, skills, and habits of mind intended by test developers are the ones that students are actually using in arriving at their item responses. Be alert for irrelevant factors that cause students to solve problems in ways different from what assessment developers intended, including ways aimed at gaming the automated system.

   d. In designing and evaluating the interface and tutorial, be particularly sensitive to the needs of all learner groups, especially those who are English language learners, students with disabilities, and computer novices. Be sure to oversample these groups when conducting usability studies and cognitive labs.

e. If automated scoring is being added to an existing assessment system, conduct a complete design review to understand how the automated scoring is likely to interact with the other system components, especially the interface and the tutorial.

2. ***Encourage vendors to base the development of automated scoring approaches on construct understanding.*** One development approach worth considering is as follows:

a. Ask domain experts to enumerate the response features that denote different levels of quality.

b. Create computable measures of those features, and identify important features that cannot be automatically evaluated.

c. Combine the computable features according to weights that the domain experts believe would reasonably reflect quality in a response.

d. Produce scores for a sample of student responses.

e. Ask a new sample of domain experts to carefully analyze each response under liberal time conditions, and accept or challenge the automated score. For any challenge, the experts should state the specific reasons. Review each challenge and reason, identifying any potential connection between the challenges and the important features that the automated scoring could not directly evaluate.

In this approach, three things should stand out. First, experts are used instead of less-authoritative operational raters. Second, experts are given sufficient time to do a detailed analysis of each response. Finally, the scoring is *not* built to predict the assignments given by raters; rather, the raters are used to verify that the automated scoring is producing meaningful results.[16]

3. ***Strengthen operational human scoring.*** In the ideal case, automated and human scoring help improve one another. For that to occur, the highest-quality human scoring feasible is needed. Online scoring, in which human judges grade responses on computer, can aid quality monitoring. Consider trying out the following:

a. Ask domain experts to carefully score a sample of responses and use their time distribution to help identify operational raters who might be scoring too quickly.

b. Provide operational raters with efficient annotation tools and ask them to mark the response features that contribute significantly to the grade they award. Periodically analyze a sample of annotations and give feedback to raters to encourage focus on the most relevant response attributes.

c. For training and qualification, force rater attention to higher-level features by using benchmarks, training samples, and qualification sets that vary primarily in those features. For example, for essays, select benchmarks that vary across score level in the quality of argumentation but that vary less in lower-level features (such as response length).

Quality monitoring can also be aided by automated scoring, especially if that scoring is developed as described in Number 2 above. Such automated monitoring might be used to identify (for closer examination) potential rater drift over time or possible errors by individual raters.

4. ***Where automated systems are modeled on human scoring, or where agreement with human scores is used as a primary validity criterion, fund studies to better understand the bases upon which humans assign scores.*** Cognitive labs with human raters, studies in which raters annotate the features of responses that they are judging, and eye-tracking methods can help inform the development of theories of human scoring. Automated approaches that are modeled on human scoring will become considerably stronger to the extent that those approaches are based on a deeper understanding of what human raters do. Without such an understanding, testing programs are replicating a *black box*, complete with any systematic biases that box might contain.[17]

5. ***Stipulate as a contractual requirement the disclosure by vendors of those automated scoring approaches selected for operational use.*** When the life chances of students, teachers, and school administrators are at stake, state officials have a responsibility to ensure that these individuals are being judged in a manner that can be explained and defended. At the same time, vendors are entitled to protect their intellectual property. To satisfy these competing needs, consider doing the following:

   a. Encourage vendors to use patent protection. In situations where both disclosure and preserving intellectual-property rights are important, patent is a solution that responsible vendors should be willing to entertain.

   b. Ask vendors to disclose what is being done and how at a level of detail sufficient for understanding and evaluation by individuals expert in the art.

   c. Conduct an evaluation of the scoring approach using experts in the art, as well as measurement specialists and content specialists in the field in question (e.g., literacy, math). In the same way as a technical advisory committee reviews an assessment program, the goal of this suggested evaluation is to assure that automated scores are being computed in a way that aligns with the construct—in short, that those scores are being generated in a manner that is likely to be valid, fair, and supportive of learning and instruction.

6. ***Require a broad base of validity evidence similar to that needed to evaluate score meaning for any assessment.*** All of the evidence categories listed may not be available before operational use, but evidence from as many categories as possible should be assembled during the field-test stage. In addition, there should be a written commitment to gather additional evidence in a rapid fashion so that the quality of automated scores can be defended and continuously improved. Evidence should be gathered from the same student populations, with the same types of test items, and from the same contexts for which generalizations about score meaning are sought. The relevance of evidence gathered from different populations, item types, and contexts should be regarded as *speculative*, at best.

The categories of evidence that should be assembled include:

a. A mapping of the scoring method to the literacy or mathematics construct of interest. That mapping should describe the response features measured by the automated scoring, how each feature is judged, and how the judgments are weighted and combined to form an item score. Each of these descriptions should be justified in terms of the specific literacy or mathematics construct, and be built upon the disclosure and expert evaluation stipulated in Number 5 above.

b. Agreement with human raters (and invariance of that agreement across population groups). For reasons discussed earlier, this agreement should ideally be computed with the scores of domain experts rendered under ideal conditions, rather than with the scores of typical raters under operational conditions. This category of evidence, as well as the next, should be evaluated within important population groups to ensure that the automated scoring functions in the same way in each group.

c. Relationships of automated scores and of human raters' scores to other relevant criteria (and invariance of those relationships across population groups).[18] Such criteria should include measures to which scores should be highly related, as well as ones to which scores should be less related. For instance, in comparison with human raters' scores on short-text math responses, the relationship of the automated scores to scores on technology-enhanced math items, math project grades, math course grades, and math accomplishments (e.g., membership in a math club) might be looked at, as well as the relationship of the human and automated scores to spelling test performance. Given what is known of the susceptibility of automated grading to spelling errors, a notably higher correlation for automated than human scores with spelling performance might suggest potential inaccuracy.

d. Susceptibility of the scoring method to manipulation and sensitivity to unusually creative responses.[19] Data should be presented to show that the automated scoring program cannot be easily gamed—that is, that undeservedly high scores can't be earned by playing on knowledge of the limitations of the scoring approach. For writing assessment, something that should be tried—because the media may attempt it if testing program staff do not—is entering a vacuous but mechanically correct, five-paragraph essay with sophisticated vocabulary that is clearly related to the question topic.[20] Also important is to ensure that especially inventive responses are not likely to receive lower scores than deserved because the quality of the reasoning or of the rhetorical devices is beyond what the machine can process.

7. ***Unless the validity evidence assembled in Number 6 justifies the sole use of automated scoring, keep well-supervised human raters in the loop.*** The strength of the evidence (on both automated and human scoring) should dictate how human raters are used. For many years, the typical practice in high-stakes testing programs has been to have each constructed-response independently rated by two human judges. For moderately accurate methods, such as automated essay scoring, most testing programs have changed standard practice by replac-

ing one of the human raters with an automated score, routing to a second human any response for which there is significant discrepancy. That practice seems consistent with the available validity evidence.

A different approach might be justified if the evidence suggested that a particular task type could be scored more robustly than in the case above. In the grading of math equations, for example, the evidence might lead to routing only a sample of responses to human raters. That sample could be chosen at random or it could be purposively selected (e.g., students whose automated scores fall outside the region predicted from their selected-response performance; students with total test scores in the region of a proficiency cut-point).

Under some circumstances, a conceivable alternative to having both a machine and human rater read every response might be to employ two different automated scoring programs. A human would still be required to intercede when the two programs disagreed. However, were this approach to work, it would still be less labor-intensive than having every response graded by a human.

The operational introduction of automated scoring might be thought of in terms of a continuum, where the most innovative (and least trustworthy) methods are always paired with well-supervised human scoring and the least innovative (but most trustworthy) methods run with only human checking of quality-control samples. All along the continuum, carefully examining the human-machine discrepancies will help identify weaknesses in both the automated and the human process, directing developers and program managers to where improvements are needed. As the automated technologies are refined and evidence amasses, they can be progressively moved up the continuum toward more independent use.

Notes

1) Randy Elliot Bennett is Norman O. Frederiksen Chair in Assessment Innovation at Educational Testing Service. The views presented in this paper are those of the author. I am grateful to Dan Eignor, Kathy Feeney, Steve Lazer, Kit Viator, and David Williamson for their helpful comments on drafts of this paper.

2) Far fewer studies have attempted to compare human and automated scoring to other validity criteria (e.g., other test scores, grades in courses requiring writing, other writing samples, self-perceptions of writing skill, writing accomplishments) or to look at the extent to which human and automated scores function similarly across population groups (e.g., white students, black students, males, females). For a rare example of the latter type of study, see Brent Bridgeman, Catherine Trapani, and Yigal Attal, "Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country" (in press), *Applied Measurement in Education*.

3) Sometimes with embarrassing, and all-too-public, results. For an example in the context of low-stakes formative assessment, see: Scott Elliott, "Nonhuman Factors," *Dayton Daily News*, May 23, 2004, retrieved November 22, 2010, from http://www.daytondailynews.com/blogs/content/shared-gen/blogs/dayton/education/entries/2004/05/23/nonhuman_factor.html.

4) For an example, see: Jana Sukkarieh and John Blackmore, c-rater: *Automatic Content Scoring for Short Constructed Responses*, Florida Artificial Intelligence Research Society (FLAIRS): Proceedings of the Twenty-Second International FLAIRS Conference, 2009, retrieved November 18, 2010, from http://www.aaai.org/ocs/index.php/FLAIRS/2009/paper/download/122/302.

5) For an example of this type of task, see: Randy Bennett, Manfred Steffen, Mark Singley, Mary Morley, and Daniel Jacquemin, "Evaluating an Automatically Scorable, Open-ended Response Type for Measuring Mathematical Reasoning in Computer-adaptive Tests," *Journal of Educational Measurement* 34 (1997): 162-176.

6) For an example of the kind of research study needed to evaluate the fairness of response entry, see: Ann Gallagher, Randy Bennett, Cara Cahalan, and Donald Rock, "Validity and Fairness in Technology-based Assessment: Detecting Construct-irrelevant Variance in an Open-ended, Computerized Mathematics Task," *Educational Assessment* 8 (2002): 27-41.

7) For a more detailed description of this class of tasks and the scoring issues, see: Randy Bennett, Mary Morley, Dennis Quardt, Donald Rock, Mark Singley, Irvin Katz, and Adisack Nhouyvanisvong, "Psychometric and Cognitive Functioning of an Under-determined Computer-based Response Type for Quantitative Reasoning," *Journal of Educational Measurement* 36 (1999): 233-252.

8) For an early example, see: Marc Sebrechts, Randy Bennett, and Donald Rock, "Agreement between Expert- System and Human Raters' Scores on Complex Constructed-response Quantitative Items," *Journal of Applied Psychology* 76 (1991): 856-862.

9) Questions like these are successfully used in intelligent tutoring programs like the Carnegie Learning Cognitive Tutor but, in that context, the same questions can be employed repeatedly, allowing the labor involved in creating automated scoring to be amortized over long periods of item reuse. See http://www.carnegielearning.com/.

10) For a detailed argument, see Randy Bennett and Isaac Bejar, "Validity and Automated Scoring: *It's Not Only the Scoring, Educational Measurement: Issues and Practice* 17(4) (1998): 9-17.

11) For essays, an average of two minutes per response would not be unusual.

12) Anecdotal evidence, at least, suggests that there are instances where raters are not using (or need not use) the intended rubric. See, for example: Todd Farley, *Making the Grades: My Misadventures in the Standardized Testing Industry* (Sausalito, CA: PoliPoint Press, 2009), 58.

13) See previous footnote. Also see: Michael Winerip, "SAT Essay Test Rewards Length and Ignores Errors," *New York Times*, May 4, 2005, retrieved November 19, 2010, from http://www.nytimes.com/2005/05/04/education/04education.html.

14) It's reasonable to ask whether such effects might be mitigated through the use of classroom formative assessment that focused on the higher-level skills of concern. However, from the research on high-stakes testing, it appears to be the case that what gets tested is taught. The principle claimed here is simply a derivative: "What gets scored is taught." If these principles weren't true in practice, approaches to automated scoring that depended solely on low-level features would be fine, as would approaches to high-stakes testing that depended solely on multiple-choice questions.

15) At least 75 percent of the weight in the automated scoring of GRE Argument and Issue essays is assigned to what are arguably such lower-level features (e.g., essay length and grammar, usage, mechanics, and style errors); 20 percent or less is assigned to "content," or more properly topical-vocabulary features related to the substance of the essay question. See Yigal Attali, Brent Bridgeman, and Catherine Trapani, "Performance of a Generic Approach in Automated Essay Scoring," *Journal of Technology, Learning, and Assessment* 10(3) (2010): 8, retrieved November 13, 2010, from http://escholarship.bc.edu/jtla/vol10/3/. In Table 1, essay length is the sum of the weights attributed to the Organization and Development features which, respectively, are the number of discourse units found in a response and the average length of the units.

16) For an application of this approach to the scoring of architectural design problems, see: Isaac Bejar, "From adaptive testing to automated scoring of architectural simulations" (1995). In Elliott Mancall & Philip Bashook (Eds.), "Assessing clinical reasoning: The oral examination and alternative methods" (Evanston, IL: American Board of Medical Specialties): 115-130. For an application to essay scoring, see Anat Ben-Simon & Randy Bennett, "Toward more substantively meaningful automated essay scoring," Journal of Technology, Learning and Assessment 6, 1 (2007), retrieved November 13, 2010, from http://escholarship.bc.edu/jtla/vol6/1/

17) The U.S. measurement-research community, by far the biggest in the world, has largely neglected research on the underlying bases of human scoring. In contrast, an exemplary program is underway in the UK, led by Cambridge Assessment. See Victoria Crisp, "Researching the Judgement Processes involved in A-level Marking," *Research Matters* 4 (2007): 13-17. Also see Irenka Suto and Jackie Greatorex, "A Cognitive Psychological Exploration of the GCSE Marking Process," *Research Matters* 2 (2006): 7-11; Irenka Suto and Jackie Greatorex, "What Goes through an Examiner's Mind? Using Verbal Protocols to Gain Insights into the GCSE Marking Process," British Educational Research Journal 34 (2) (2008): 213-233; and Irenka Suto & Jackie Greatorex, "A Quantitative Analysis of Cognitive Strategy Usage in the Marking of Two GCSE Examinations," Assessment in Education: Principles, Policies and Practice 15 (1) (2008): 73-90.

18) For a recent example, see: Attali, Bridgeman, and Trapani, "Performance of a Generic Approach in Automated Essay Scoring," Table 3, p. 11, retrieved November 13, 2010, from http://escholarship.bc.edu/jtla/vol10/3/. For an exemplary investigation, see Donald Powers, Jill Burstein, Martin Chodorow, Mary Fowles, and Karen Kukich. *Comparing the Validity of Automated and Human Essay Scoring* (RR-00-10), (Princeton, NJ: Educational Testing Service, 2000), retrieved November 13, 2010 from http://www.ets.org/research/policy_research_reports/rr-00-10

19) For an exemplary study, see Donald Powers, Jill Burstein, Martin Chodorow, Mary Fowles, and Karen Kukich, "Stumping E-Rater: Challenging the Validity of Automated Essay Scoring," *Computers in Human Behavior* 18 (2002): 103-134.

20) See Elliott, "Nonhuman Factors."