Automated Scoring for the Assessment of Common Core Standards

David M. Williamson, Senior Research Director, Applied Research and Development, ETS

Randy E. Bennett, Frederiksen Chair, Assessment Innovation, ETS

Stephen Lazer, Vice President, Student and Teacher Assessments, ETS

Jared Bernstein, Founder of Ordinate (now Knowledge Technologies),
Consultant to Knowledge Technologies, Assessment and Information Group

Peter W. Foltz, Vice President for Product Development, Pearson Knowledge Technologies, Pearson

Thomas K. Landauer, Vice President for Research, Pearson Knowledge Technologies, Pearson

David P. Rubin, Senior Vice President, Engineering, Pearson Knowledge Technologies, Pearson

Walter D. Way, Senior Vice President, Psychometric and Research Services, Pearson

Kevin Sweeney, Executive Director, Psychometrics, The College Board

July 2010

Introduction

This paper discusses automated scoring as a means for helping to achieve valid and efficient measurement of abilities that are best measured by constructed-response (CR) items.

Multiple consortia have formed to pursue the development and use of Race to the Top assessments. Current visions for these assessments call for use of CR items that require students to produce written or spoken responses, or to construct solutions to mathematical problems. Thus, the consortia will face the challenge of scoring responses from large numbers of students, with each student potentially responding to numerous CR items across English language arts (ELA) and mathematics assessments.

One possible mechanism for scoring the CR items is with human graders. However, years of research and practical experience with human graders reveal a number of challenges. There is a substantial and expensive logistical effort in supporting human scoring through recruiting, training, monitoring, and paying human graders. The process also takes considerable time to accomplish and can make it difficult to report scores quickly. Finally, even under the best conditions, there can be limits to the objectivity and consistency of human scores, despite using multiple human graders for each response.

Human scoring is not the only option for scoring CR items. Recent advances in artificial intelligence and computing technology, paired with research over several decades, have yielded a number of systems in which computers assign scores to CR items. Both the availability and acceptance of automated scoring systems have grown substantially in recent years, with year-over-year expansion of operational use in both high- and low-stakes assessments. The growth in the use of automated scoring is due to the ability of such systems to produce scores more quickly and at a lower cost than human scoring. In addition, because automated scoring systems are consistent in how they produce scores, they support longitudinal analysis of performance trends and equity in scoring. Some applications of these systems provide detailed feedback on aspects of performance. This information can be provided to individual test takers or can be aggregated for use at the classroom, school, district, or state level. However, these advantages come with certain limitations, and care must be taken to use automated scoring systems appropriately.

The purpose of this paper is to share information about the current capabilities of automated scoring systems, discuss how these systems can help meet the goals of Common Core assessments, and describe prudent steps for evaluating and deploying automated scoring. The paper is organized into three sections:

- Using state-of-the-art automated scoring
- Ensuring the measurement quality of automated scoring implementation
- Deploying automated scoring

It is hoped that this paper will support sound planning and decision making related to the use of automated scoring systems in Common Core assessments.

Using State-of-the-Art Automated Scoring

There are a variety of automated systems available for scoring multiple kinds of CR items, including essays, spoken responses, short text answers to content questions, and textual, numeric, or graphical responses to mathematical questions. With one exception, all of the scoring systems for task types described below require delivery of items and entry of responses via computer and are not designed to score handwritten responses. The exception is spoken responses, which can be provided via telephone or computer. The availability and maturity of these capabilities vary based on the types of items they are designed to address. This section will provide a brief overview of the state of the art of automated scoring for these response types.

Mathematical Responses

In the area of mathematics, the performance of automated scoring systems is typically quite robust when the response format is constrained. The types of mathematics item responses that can be scored by automated systems include mathematical equations or expressions, two-dimensional geometric figures, linear, broken-line or curvilinear plots, bar graphs, and numeric entry. The field has experienced at least eight years of advances in these systems since they were first deployed in consequential statewide assessments, and it is reasonable to expect these systems to perform with high accuracy. This enables the use of these systems without additional oversight by human raters. Automatic scoring of freehand graphic responses and handwritten expressions achieves lower accuracies. For the more constrained response types, the most notable limitation is that automated scoring assumes computer test delivery and data capture, which in turn may require an equation editor or graphing interface that students can use comfortably.

Essays

The most widely available automated scoring systems are those for scoring essays, although most have been deployed in learning environments and only a few have been used in high-stakes assessment. All automated scoring systems for essays rely on the essay being typed by the examinee; they do not score handwritten responses. Those that have been used in high-stakes assessment have demonstrated levels of agreement between automated scores and human scores that are comparable to agreement rates among human raters. However, the agreement between human raters may be lower than desired. Thus, agreement with human scores may not always be a sufficient accomplishment.

Automated essay scoring systems do not measure all of the dimensions considered important in academic instruction. Most automated scoring components target aspects of grammar, usage, mechanics, spelling, and vocabulary. Therefore, they are generally well-positioned to score essays that are intended to measure text-production skills. Many current systems also evaluate the semantic content of essays, their relevance to the prompt, and aspects of organization and flow. Assessment of creativity, poetry, irony, or other more artistic uses of writing is beyond such systems. They also are not good at assessing rhetorical voice, the logic of an argument, the extent to which particular concepts are accurately described, or whether specific ideas presented in the essay are well founded. Some of these limitations arise from the fact that human scoring of complex processes like essay writing depend, in part, on "holistic" judgments involving multivariate and highly interacting factors. This is reflected in the common use of holistic judgments in human essay scoring, where they may be more reliable than combinations of analytic scores.

Short Content Responses

Short text CR tasks are designed primarily to measure student content knowledge and skills, rather than writing ability. Short CR tasks require a student to respond with short text demonstrating his or her understanding of key concepts in a domain. For example, such a task in ELA might ask the student to identify the contrasting goals of a protagonist and an antagonist in a passage. Automated scoring systems can evaluate the textual responses to determine whether the goals of the two opposing characters are appropriately represented in the response.

One challenge associated with such systems is to develop items with definitive correct answers that the automated scoring system can verify. If the items call for opinions or other unverifiable discussion, the expected response set becomes less certain and more difficult for the automated scoring system to handle. Thus, a variety of factors influence the success of these systems for scoring, including the number of potential concepts that could be generated in a response, the variety of ways in which these concepts might be expressed, and/or the degree to which there is a clear distinction between correct and incorrect representations of the concept, among others.

There is less operational experience with automated scoring systems of short texts than with essay scoring. Also, it is often the case that human raters score short content responses at considerably higher agreement rates than they do essay responses, which creates a higher standard for automated scoring methods to attain. Thus, there is still progress to be made in developing a clear understanding of which kinds of items will work best with automated scoring in a short-response format.

Spoken Responses

Scoring systems also have been deployed that score spoken responses for the quality of English-language production. There are two general classes of automated scoring systems for speech: those that are designed primarily for scoring predictable speech and those designed to score unpredictable speech.

Tasks that elicit relatively predictable speech include read-aloud tasks, short answers to specific questions (e.g., "How long until the plane arrives at the airport?"), repeating spoken prompts, paraphrasing spoken passages, and tasks in which the student describes a picture or other situation in which there is a specific range of possible correct responses. When the automated scoring systems are properly trained and calibrated on appropriate samples, the quality of the automated scoring for predictable speech is generally sufficient for use in consequential assessment. These assessments are typically scored primarily for the intelligibility and linguistic forms of spoken English, but they also can represent domain content to the extent that the desired responses are provided to specific questions (e.g., any form of correct answer to the specific question of when the plane arrives at the airport). The general public may be familiar with the use of automated systems for predictable speech through exposure to telephone voice recognition systems that are widely used by corporations to respond to customer questions. However, the technology used in scoring language skills from predictable-response tasks is based on more accurate methods that have been used in high-stakes assessment and have demonstrated levels of agreement between automated scores and human scores that are comparable to agreement rates among human raters.

A task designed to elicit unpredictable speech might include a prompt such as, "If you could travel to any place in the world, where would it be and why?" In this case, the expected response set is much less predictable in both topic and specific forms that might be used. As a result of the recent emergence of this type of automated scoring, fewer applications of automated scoring systems for relatively unpredictable speech have been evaluated and fielded. Automated systems that evaluate unpredictable speech have not yet been used for high-stakes testing purposes.

Simulation-Based Tasks

A final class of CR task types is computer-based *simulations*, which are intended to replicate important features of real-world situations so that the constructs typically called upon in those situations can be evoked and measured. Simulation-based assessments are common in a variety of professional settings, such as in the assessment of pilots, physicians, and architects. The strengths of this approach lie in the flexibility and range of task types that are possible in a simulation-based environment, while still being entirely scoreable by machine. However, the high degree of effort involved in designing tasks, developing reliable human-based scoring approaches, and automating the scoring process is clearly a challenge.

Ensuring the Measurement Quality of Automated Scoring Implementation

Despite advances in the state of the art of automated scoring, there is much that is not yet known about the performance of these systems. The following is intended to help define how to answer the question, "How do you know automated scoring works effectively?" For convenience, this overview is presented as a practitioner's "checklist" for evaluating the measurement characteristics of automated scoring. The intent is to provide a tool for consortia to use in requesting evidence from vendors that the automated scoring meets expectations for performance in Common Core assessments.

✔ *Automated scores are consistent with the scores from expert human graders.* In order to be an appropriate substitute for human scores in Common Core assessments, the automated scores should produce results similar to those of a human rater. This similarity is typically demonstrated through statistical measures of agreement between automated and human scores, such as correlations and weighted kappa (rather than percent agreement, which may overestimate the agreement rate between automated and human scores). Ideally, the automated scoring will demonstrate a level of agreement with human scores comparable to the agreement between two or more independent human raters. Other measurement advantages of automated scoring could offset a performance deficit in this regard. For example, if the use of automated scoring permits adding items to the assessment that otherwise would not be feasible because of time and/or cost limitations associated with human scoring, measurement could be improved even if the automated scoring for each item was slightly less consistent than human scoring. In any event, it is important that the overall distribution of automated scores also approximates that of the human score distribution so that automated scores are not systematically higher or lower, or more or less dispersed, than their human counterparts. This evaluation should be conducted for each individual CR item rather than in the aggregate, as aggregate analyses can mask potential anomalies in the performance of a particular item. It also may be useful to check that all elemental constructs (and reported subscores) are working well in relation to operational human scoring performance.

✔ *The way automated scores are produced is understandable and substantively meaningful*. Striving to achieve automated scoring results that are similar to human scoring is important. But so, too, is applying an automated scoring methodology that allows for the review, understanding, and evaluation of the scoring mechanisms as relevant and valued by experts in the content domain being measured. To the extent that automated scores can be based on meaningful and construct-relevant characteristics of the response, positive "washback" effects for education can be encouraged, so that the actions teachers take in preparing students to succeed on Common Core standards assessments are the same steps they would apply in the domain of practice. That is, the influence of automated scoring (and its influence on task design) should not distort instruction, at least not to any greater degree than is the case for human-scored CR tasks. Transparency also can allow users to evaluate components of a scoring system, including the potential for scoring features that are predictors of human scores but unrelated to the constructs being tested. For example, in scoring essay assessments, the relationship between the number of words in an essay and a human score is almost as strong as the relationship between two human scores. However, it would be undesirable to have scores generated primarily from word count because such generation might encourage the student to maximize the number of words at the expense of other, more valued aspects of writing.

- ✔ *Automated scores are fair.* It is critical that automated scoring be equitable for persons from diverse groups. Research has found that in some instances, otherwise accurate automated scoring systems have shown a difference in the agreement between human and automated scores based on demographic group membership. It is not yet known whether this difference represents a concern with scores from human graders, automated systems, or both. To the extent that an automated system may be biased, that bias may be consistent and pervasive, potentially accumulating to impact the test scores.

- ✔ *Automated scores have been validated against external measures in the same way as is done with human scoring.* The comparison between automated and human scores on the same item is part of the evaluation of automated scoring quality. But it is equally important to validate against measures that are external to the specific items in the test as should normally be done with human CR scoring. Examples of relevant external criteria include scores on other test sections, grades in relevant academic classes, scores on the same test section on alternate occasions, and scores on specially designed external measures of the construct of interest. Ideally, automated scores should be more strongly related to the targeted constructs and less strongly related to constructs different from what the automated scores are intended to measure. The pattern of relationships with external criteria should be compared to the pattern found for human raters with those same criteria to determine whether that pattern implies levels of validity for the automated system that are greater than, similar to, or less than the validity for human scores.

- ✔ *The impact of automated scoring on reported scores is understood.* There are a number of different automated scoring systems, and in a single assessment, multiple systems might be applied across multiple item types. Sometimes, small differences at the item level can accumulate into large differences at the test level. Therefore, even if, item by item, the automated scoring appears to perform well, an evaluation at the test level may reveal notable differences between automated and human scores.

With respect to the above list, it is worth emphasizing that the basis for checking these criteria should be from results on the Common Core assessment (or pilot test), rather than results from some other assessment and/or population of examinees. Therefore, being able to check these boxes requires research during the design and development phase of Common Core standards assessment. In this process, it is reasonable to expect that some aspects of computer and human scoring will work as intended and others will not, requiring modification in the task design or the assessment design.
In addition, the use of automated scoring often requires a calibration and evaluation process for these capabilities that must be planned for in the design and delivery of Common Core assessments. Furthermore, it is prudent to plan for the possibility that there may be multiple ways in which the automated scoring capabilities might be used in Common Core assessments. The following section outlines some of the potential models for implementation of automated scoring and the implications for Common Core assessment design.

Deploying Automated Scoring

There are a number of ways in which automated scoring might be deployed. The decision about how to deploy is influenced by two aspects of the automated scoring system: measurement characteristics and speed. The measurement characteristics of automated scoring are typically a function of the automated scoring system, the type of task it is applied to, the population, and the purpose for which the system and scores are to be used. In contrast, the speed of response is a combination of the time required for the computer to score the response, the speed of the connection between the administration terminal to the scoring server, and the capacity of scoring servers available to meet scoring demand at any given time. As such, the response time can be reduced to some extent through greater investment in the hardware infrastructure for scoring.

One way to deploy automated scoring is as the sole score for an item, with no human oversight or intervention. This model is typical when the measurement characteristics of the automated scores are sufficient for the purposes of the assessment. The item types that most lend themselves to this use of automated scores are those for which the expected set of responses is relatively predictable, such as some mathematics items, some types of simulation-based tasks, some short content responses, and some types of spoken responses. Use of automated scoring alone for other task types, including essays, is common in low-stakes assessment but less common for high-stakes assessment. If the speed of scoring is very high for such systems, the item can be used like any other in computerized assessment, including in computerized adaptive tests. However, if the scoring is not immediate it may complicate, but not preclude, aspects of an adaptive testing design.

An alternate way to deploy automated scoring is as one of two scores for a response, with the other score being provided by a human rater. This model is common primarily for essay scoring, with some short content responses and some types of spoken responses also potentially lending themselves to this approach. The speed of automated scoring is less of a factor in this model, since the final score would not be determined until after the human score was also obtained.

Conclusion

This paper has provided an overview of considerations regarding the use of automated scoring capabilities in Common Core assessments, along with elements in a process for evaluating such scoring. It highlights some of the promising potential of automated scoring, as well as some of the current challenges and limitations of these capabilities in consequential assessment. These current limitations and challenges are the subject of active research on the part of multiple institutions that may result in enhancements to the state of the art of automated scoring within the next few years. There are a number of different kinds of CR items that can be scored well by automated systems. Understanding the current state of the art of these systems, and the recommended practices for achieving the measurement characteristics that are appropriate for the Common Core assessments, should help consortia make prudent decisions about both the goals for use of automated scoring and the process by which those systems will be evaluated and deployed. In so doing, the consortia will be in a better position to leverage the potential of automated scoring in the best way for Common Core assessments.