Validity and Automated Essay Scoring:
A Summary Paper

Randy Elliot Bennett
and
Mo Zhang

Educational Testing Service
Princeton, NJ 08541
rbennett@ets.org
mzhang@ets.org

Keywords: automated scoring, performance assessment, writing assessment

Paper presented at the annual meeting of the International Association for Educational Assessment, Tel Aviv (Israel), October, 2013. Any opinions expressed herein are those of the author(s) and not necessarily of Educational Testing Service.

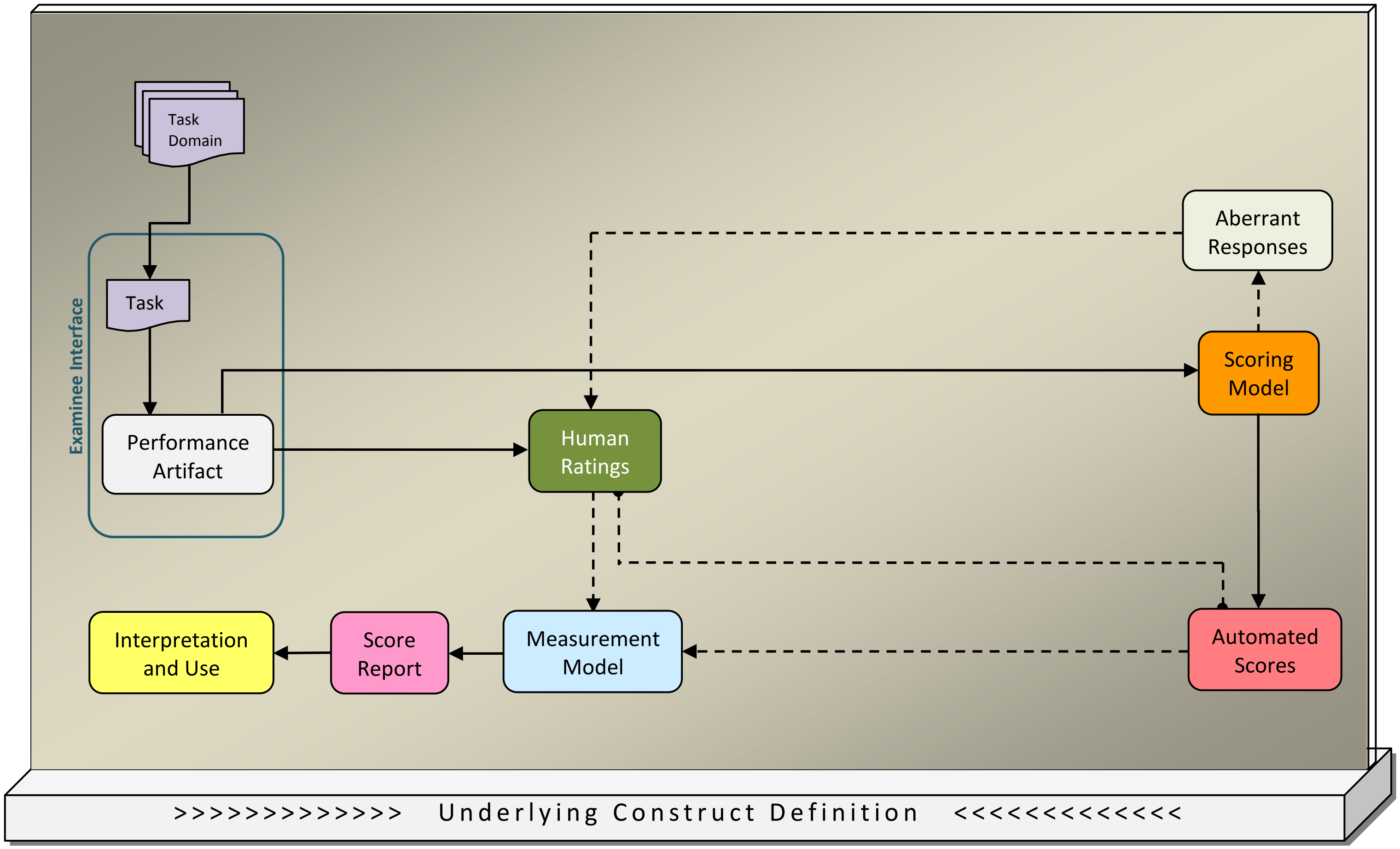
In the United States, automated essay scoring is used in many large-scale testing programs. It can be found in postsecondary admissions tests (e.g., GRE revised General Test, TOEFL iBT, Graduate Management Admission Test, Pearson Test of English), in professional licensure and certification (e.g., Uniform Certified Public Accountant Examination), in college placement (e.g., ACT COMPASS, College Board ACCUPLACER), and in state assessment (e.g., Wyoming Direct Writing Assessment).

This paper summarizes the automated essay scoring process, the process of scoring model generation, and some of the evidentiary sources and the questions underlying the validity argument for automated essay scoring as they might relate to measuring writing proficiency in computer-based tests like those listed above. Greater detail can be found in Bennett and Zhang (in press).

Automated Essay Scoring in Operational Use

Figure 1 gives a high-level overview of automated essay scoring in operational use. The process begins with a task or, in this case, an essay prompt, being drawn from a pool of similar prompts generated from an underlying construct definition. The prompt is presented to an examinee via a computer interface which also provides a mechanism for that examinee to respond, allowing him or her to produce a performance artifact, or essay. In most consequential testing programs, that essay is routed to the automated essay scoring program as well as to a human rater. Both the rater and the program produce a score, the rater using a rubric aligned to the construct definition and exemplars (i.e., representative essays for each score level), while the program uses a “scoring model” (described below). As part of the automated scoring process, essays with aberrant characteristics (e.g., too few words) may be filtered and further examined by human graders for final judgment. Human and machine scores are next combined, possibly with other item responses (e.g., other written responses, responses to multiple-choice questions), through a measurement model. That model produces a reported score which is given some interpretation and use.

Figure 1. Automated essay scoring in operational use. From R. E. Bennett and M. Zhang (in press).



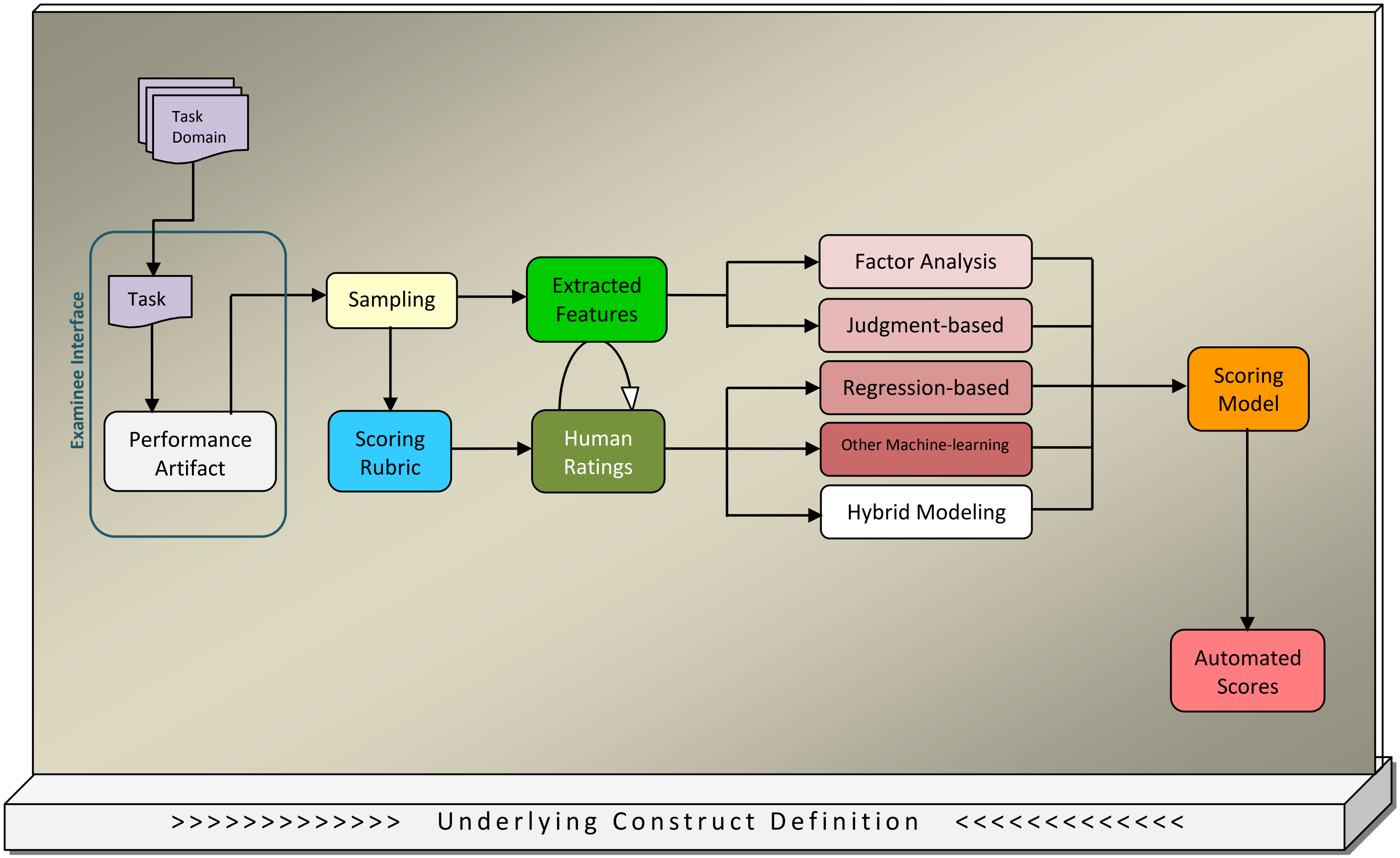
Note. Dashed lines indicate that a connection is optional.

Generating the Scoring Model

As noted above, automated scores are produced via a scoring model. Figure 2 gives a high-level overview of the model creation process. That process begins with the selection of a sample of essay responses from the examinee population. That sample is scored by human raters, ideally two or more. Separately, the automated system extracts text features from each response. Those features may include such elements as word length; essay length; sentence variety; accuracy in grammar, usage, mechanics, and style; word difficulty; overlap of words or word meanings among adjacent text segments; and the similarity of words to other human-scored essays on the same topic. The extracted features are next assembled into a scoring model, an algorithm that combines the features and weights them. Whereas many such assembly methods are possible, the most common approach is probably that of regressing the human scores (linearly or logistically) onto a set of the automatically extracted features. The resulting model can then be used to produce scores for as-yet-unseen responses. These scores are essentially predictions of the scores that human raters would assign had they done the grading themselves.

In addition to regression approaches, methods that also use statistical optimization to select and combine features so as to best predict human ratings include such other machine learning techniques as support-vector machines and decision trees. Optimization techniques (including step-wise regression) may capitalize on features that are highly related to human scores but that do not reflect the construct fully, essay length being a commonly cited example. Other than prediction-oriented approaches, automated scores may be produced via methods not based on human ratings, such as using a committee of content experts to select text features and their relative influence, or using factor analysis of selected text features to generate loadings for use as scoring-model weights. Finally, hybrids of these various methods might be adopted.

Figure 2. Deriving automated essay scoring models. From R. E. Bennett and M. Zhang (in press).



Note. The circular arrow between "human ratings" and "extracted features" denotes the potential joint use of human ratings and features in producing the scoring model.

Evidentiary Sources and Questions Underlying the Validity Argument for Automated Essay Scoring

Without doubt, the most common means of "validating" automated scores has been to compare them to the scores of human raters (Attali, 2013; Bridgeman, 2013), and most often, operational human raters. However, as has been noted elsewhere, agreement with operational human rating represents a limited source of validity evidence (Bennett, 2011; Bennett & Bejar, 1998; Zhang, 2013). Most of the sources of evidence that might need to be considered in formulating a more complete validity argument can be inferred from Figure 2. Several of those sources concern the meaning of the human ratings themselves (Bejar, 2012), including their interaction with the examinee response process, rubric, and rater cognitive process. For any number of reasons, the processes in which examinees engage when responding might not align with the construct definition, which would in turn cause the human ratings to be a contaminated indicator. Such an effect could occur, for example, if the response interface was unfamiliar, making the ratings as much an indicator of interface familiarity as of writing proficiency. A second source of evidence is the rubric itself, in particular the extent to which it fully captures the construct definition. If it does not, the human scores may under-represent the competency of interest. Assuming that the rubric is fully relevant, a third source of evidence relates to the extent to which rater scoring processes align with the rubric. Short-cuts, like judging the response on the basis of the opening and closing passages, or eye-balling the length, may lead to an erroneous validation criterion. Fourth is the extent to which raters agree among themselves in judging the same set of performances. Frequent, large discrepancies signal an undependable criterion. A fifth concern regards the degree to which raters appropriately grade responses that are atypically creative or that attempt to game their way into higher scores than deserved. Sixth, do human ratings on one task predict performance on other tasks from the task universe? An absence of a pattern of positive relationships would suggest a failure to realize a coherent writing construct through human rating. A seventh source of evidence can be found in the degree to which human ratings are more strongly associated with other indicators of writing than with related, but different, constructs (e.g., editing skill). A final source is the extent to which each of the above characteristics holds across demographic groups considered important to testing program directors or their constituents. Noticeable variation might constitute bias, which should not be modeled in automated scoring.

For most testing programs, some (perhaps most) of the questions implied above will be unanswered, making the use of human rating as the sole (or even primary) validation criterion untenable. Evidentiary sources and questions that go beyond the meaning of human rating can also be inferred from Figure 2, for which we give some examples. A fundamental issue is whether the scoring model was based on an appropriate sample from the target population. As Zhang & Qian (2013) have shown, the choice of a particular sample of responses can make a difference in the meaning of automated scores. A second source of evidence concerns the features themselves, including whether they are related to one another empirically in theoretically meaningful ways, and whether they and their weighting fully capture the rubric (and construct definition). Third is the extent of agreement with human ratings, ideally, taken across multiple human experts, thereby making for a more reliable criterion. A fourth issue is how effectively, relative to human raters, automated scoring systems identify aberrant or unusual responses that warrant additional human evaluation. Fifth, do automated scores on one task predict performance on other tasks from the task universe, and do they do so at least as well as expert human raters do? A sixth source of evidence is the relationship, compared with human rating, of automated scores to other indicators of writing, as well as to related but different constructs. Seventh, do the above characteristics hold across important population groups and are the characteristics similar in their invariance patterns to human-rater patterns? A final source of evidence is the degree to which students and teachers change their learning and instructional behavior in positive or negative ways because of the use of automated scoring. Such changes, particularly ones that center on increasing scores without necessarily improving writing proficiency, can serve to undermine the validity argument.

Conclusion

In this summary paper, we described the automated essay scoring and model generation processes. We then used those process descriptions to describe the validation effort in terms of its components, each of which implies an evidentiary source and one or more research questions that might underlie a validity argument for automated scoring. Of note is that Figures 1 and 2 also suggest that many of those components, although distinct, are interdependent. For example, human ratings are partly dependent on the adequacy of the scoring rubric; the construct coverage of the automated scoring model depends on the nature of the text features used in the model; and the characteristics of the scores produced by the measurement model depend on those of the human and automated scores put into that model. As Kane (2006) has pointed out, a validity argument consisting of a chain of dependencies is only as strong as its weakest dependency. Therefore, the validation of automated essay scoring should be viewed as an analytical activity aimed at an integrated holistic judgment, rather than as evidence-seeking from disconnected activities.

The descriptions presented in this paper can serve as mental models of the processes associated with automated essay scoring. In addition, those descriptions can be used as a high-level, practical guide for testing programs to gather and evaluate the evidence needed to justify the interpretation and use of such scores.

References

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation: Current applications and new directions (pp. 181-198). New York: Routledge.

Bejar, I. I. (2012). Rater Cognition: Implications for Validity. Educational Measurement: Issues and Practice, 31(3), 2-9.

Bennett, R. E. (2011). Automated scoring of constructed-response literacy and mathematics items. Washington, DC: Arabella Philantropic Advisors. Available: http://www.ets.org/s/k12/pdf/k12_commonassess_automated_scoring_math.pdf

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., & Zhang, M. (in press). Validity and automated scoring. In F. Drasgow (Ed.), Technology in testing: measurement issues. Washington, DC: National Council on Measurement in Education.

Bridgeman, B. (2013). Human ratings and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation: Current applications and new directions (pp. 221-232). New York: Routledge.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Zhang, M. (2013). Contrasting automated and human scoring of essays (Report Number: RDC-21). Princeton, NJ: Educational Testing Service.

Zhang, M., & Qian, J. (2013). Implementing sampling optimization for automated scoring model calibration. Paper presented at the 78th annual meeting of the Psychometric Society, Arnhem, the Netherlands.