MEASUREMENT INCORPORATED

EXCELLENCE IN ASSESSMENT™

A Gentle Introduction to Automated Scoring

Corey Palermo, Ph.D.
Measurement Incorporated

Writing skills are essential for success in school, career, and society. Unfortunately, students in the United States are not performing well in writing. For example, on the most recent National Assessment of Educational Progress (NAEP) writing assessment, nearly 75% of 8th and 11th grade students were *Basic* or *Below Basic* writers (National Center for Education Statistics, 2012). This general lack of writing proficiency carries over to the workplace, where it costs employers as much as $3.1 billion annually to remedy writing deficiencies (National Commission on Writing, 2004).

Like learning any complex skill, writing development is largely a function of practice. However, teachers juggle myriad instructional responsibilities and often have to limit students' writing opportunities to make grading manageable (Kellogg & Whiteford, 2009). Without extensive practice and regular feedback, students' writing performance will not substantively improve.

As computer-based summative assessment becomes ubiquitous, so does the opportunity to report assessment results in hours or days rather than weeks or months. However, performance tasks have historically been scored by professional raters or teachers, a time-consuming and costly process. Why not omit performance tasks from summative assessments? As opposed to machine-scored selected-response items, essays and other performance assessments tend to be more authentic, are more appropriate for measuring deeper learning outcomes, and positively impact teaching and learning (Darling-Hammond & Adamson, 2010, 2013).

Automated scoring solutions have been applied successfully to address all of these issues.

How it Works

Automated-scoring engines do not "read" student responses and assign scores in the way that people do. Rather, they match various characteristics of responses with the scores assigned by expert raters and use these relations to predict scores for new responses. Automated engines can score a variety of response types, ranging from short answers of a few words or sentences to extended essays.

Automated-scoring engines often use machine learning to determine how features of responses relate to scores assigned by expert raters. This is similar to the machine learning Netflix applies to subscribers' viewing habit data and Amazon applies to customers' page-view and purchase data to make recommendations. The main difference is that rather than predicting a collection of shows or products best matched with the user, automated scoring-engines predict the single, most appropriate score for each response.

The process used by Measurement Incorporated's scoring engine, Project Essay Grade (PEG), works like this.[1] First, a large, representative sample (1500–2000) of student responses for each item is scored by raters according to the scoring criteria[2]. Professional raters do not always agree on the score to assign a student's response, so any disagreement between raters is resolved so there is a high degree of confidence in the scores. Next, these training responses are analyzed for hundreds of linguistic features that are used to describe them in mathematical terms. For example, one relatively simple feature might address the number or severity of grammatical errors in an essay. A more complex feature might describe a pattern that represents clarity of communication. Once these features have been identified, they are associated with the scores assigned by raters. Those features of responses that do the best job of "explaining" the raters' scores are combined into a model. Finally, to verify that the new model scores responses accurately, it is validated using a small sample (300) of responses not included during training. Various measures of agreement between the engine's and raters' scores are evaluated, from simple (percent agree) to complex (Quadratic Weighted Kappa, a measure that corrects for chance agreement). A model is generally only used operationally if it meets or exceeds the accuracy of multiple raters.

Criticisms of Automated Scoring

A few concerns related to automated scoring have been raised by the field. One is that students might trick automated-scoring technologies and receive undeserved high scores, for example, by producing essays with complex syntax and vocabulary but inaccurate content (Bejar, Flor, Futagi, & Ramineni, 2014; Higgins & Heilman, 2014). While researchers have gone as far as simulating the effects of gaming strategies such as lexical substitutions, in reality a student would need to be a sophisticated writer to be capable of producing a syntactically complex but factually inaccurate essay. Also, modern scoring models rarely contain simple, linear relations among features and scores that a student might intuit and manipulate. In sum, gaming attempts are unlikely to present a significant threat to the validity of automated scores. Advancements in automated-scoring technologies are also expected to improve gaming detection. A related concern is that automated-scoring technologies emphasize surface features of text such as length. Here it is important to remember that automated scoring models the decisions of professional raters, so the extent to which surface features are emphasized is a function of the value people assign these features.

Another concern is that the automated-scoring models are a "black box," in that the scoring models and weighting of text features lack transparency. In fact, rater scoring is subject to the same limitations. Though published rubrics are an attempt to explicate the rating process, raters apply scoring criteria individually and vary in this application both within and between raters (Wilson & Andrada, 2016).

A final concern is that automated scoring compromises the social nature of writing and risks replacing teacher feedback. In practice, teachers play a critical role in helping students interpret and incorporate automated feedback (Palermo, 2017). Wilson and Czik (2016) provided evidence that teachers do not give less feedback when their students receive automated feedback. In their study, teachers in fact gave proportionally more feedback on higher-level writing skills when automated feedback was available.

[1] Most automated-scoring engines use a similar process.
[2] Scoring criteria include but are not limited to rubrics and anchor responses, or exemplars of each score point.

Benefits for Formative Assessment

Automated scoring offers many benefits in the context of assessment for learning. For one, it allows students to write and revise more by relieving the bottleneck associated with teacher grading and feedback. This accelerates the practice-feedback cycle essential for students to develop writing proficiency.

Students benefit most from specific, detailed writing feedback that is provided immediately and addresses both surface-level and content features of writing (Patthey-Chavez, Matsumura, & Valdes, 2004). Thus a second benefit is the wealth of feedback that an automated-scoring engine can produce for each response. Students who use the formative assessment program PEG Writing, for example, receive a score report that includes spelling and grammar feedback, feedback for each of six traits of writing quality, and also holistic and trait scores for each essay composed or revised.

An additional benefit of automated feedback is that it augments teacher feedback, allowing teachers to focus feedback efforts on higher-order aspects of writing. In a web-based learning environment, automated feedback can be linked to learning opportunities. For instance, PEG Writing recommends suitable interactive lessons based on students' writing quality.

A final benefit is that as part of a formative assessment program, automated scoring and feedback support deliberate practice. Sustained deliberate practice requires appropriate practice opportunities, motivation for task engagement, effortful activity to improve performance, immediate feedback based on performance, and extensive practice (Ericsson, Krampe, & Tesch-Römer, 1993). This is difficult if not impossible to achieve in classrooms where students are not writing regularly. An automated writing-evaluation program can support deliberate practice through providing writing quality feedback, evidence of growth, and efficiency; and also through supporting teachers' writing instruction and students' intrinsic motivation (Palermo, 2017). In formative contexts, the ultimate goal of automated scoring is to improve the quality of student writing.

Benefits for Summative Assessment

Automated scoring also offers a variety of benefits for assessment of learning. One benefit is that it is much faster than scoring by teachers or professional raters; once models have been generated, responses can be scored in seconds. This allows assessment results to be available to stakeholders very rapidly. A second benefit is that automated scoring tends to be as accurate or more accurate than multiple professional raters. Furthermore, automated-scoring engines are perfectly reliable in ways that raters are not—an automated-scoring engine will assign the same score to a response every time.

The cost-effectiveness of automated scoring is another benefit for summative assessment. Policymakers often weigh the costs of local scoring by teachers (potentially inexpensive but requires lots of training and burdens teachers) against those associated with central scoring (rigorous training, qualification, and monitoring of raters but more expensive). Automated scoring typically offers cost savings compared to central scoring while meeting or exceeding professional-rater quality standards.

Due to the high stakes associated with summative assessment, relatively small investments in testing[3] have a profound impact on instruction (Darling-Hammond & Adamson, 2013). Possibly the greatest benefit of automated scoring in the summative space is the potential it offers to support teaching and learning. Unfortunately, summative tests have historically measured low-level knowledge and skills due to a reliance on multiple-choice items. While assessments aligned to next-generation standards[4] place less of an emphasis on items that only require recognition, concerns about handscoring time and costs continue to limit the use of performance tasks on summative tests. By scoring such tasks efficiently, automated scoring can support assessments of deeper learning, which in turn can encourage positive shifts to instruction and influence student learning (Darling-Hammond & Adamson, 2010).

Automated Scoring Research

Much of the early research investigating automated scoring examined the validity of scores predicted by automated-scoring engines, comparing automated scoring to the "gold standard" of multiple professional raters. This line of research has generally concluded that the more advanced automated-scoring engines are capable of matching the scoring accuracy of multiple raters (e.g., Shermis, 2014).

Recent research has focused on automated scoring as a tool to support K–12 teaching and learning. Researchers found that students who used an automated writing-evaluation program wrote and revised more and were more motivated to write (Grimes & Warschauer, 2010; Warschauer & Grimes, 2008). There is modest evidence that automated feedback positively impacts students' writing quality (Morphy & Graham, 2012; Stevenson & Phakiti, 2014). Teachers reported that automated writing evaluation reduced their grading burden and supported individualized instruction (Warschauer & Grimes, 2008).

PEG Research

As a mature automated-scoring engine, PEG has been the subject of a number of studies. Results have shown PEG Writing associated with improvements in writing quality across revisions (Wilson & Andrada, 2016; Wilson, J., Olinghouse N. G., & Andrada, 2014) and across prompts (Palermo, 2017). Wilson (2017a) found that students with disabilities who used PEG Writing showed more rapid growth in writing quality than their peers due to accelerated growth in higher-level writing skills. PEG is capable of accurately identifying struggling writers who are at risk of failing state writing assessments (Wilson, Olinghouse, McCoach, Andrada, & Santangelo, 2016). Finally, Wilson (2017b, c) found that students who used PEG during the year scored higher on ELA and writing summative tests. Here PEG use explained around 10% of the variation in performance after controlling for prior achievement and demographics.

This paper has provided an overview of automated scoring, including a description of how the technology works, criticisms, benefits for formative and summative assessment, and a brief summary of relevant literature. It is hoped that this overview provides the reader with a basic understanding of automated scoring and the benefits associated with applications in various assessment contexts.

[3] On average, less than two-tenths of 1% of K–12 per-pupil spending is allocated to summative assessment (Darling-Hammond & Adamson, 2013).

[4] Examples include the Common Core State Standards (CCSS) and Next Generation Science Standards (NGSS).

References

Bejar, I. I., Flor, M., Futagi, Y., & Ramineni, C. (2014). On the vulnerability of automated scoring to construct-irrelevant response strategies (CIRS): An illustration. *Assessing Writing*, *22*, 48–59.

Darling-Hammond, L. & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., & Adamson, F. (2013). *Developing assessments of deeper learning: The costs and benefits of using tests that help students learn*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, *100*(3), 363–406.

Grimes, D. & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*, *8*(6), 1–44. Retrieved from https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1625/1469

Higgins, D., & Heilman, M. (2014). Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement: Issues and Practice*, *33*(3), 36–46.

Kellogg, R. T., & Whiteford, A. P. (2009). Training advanced writing skills: The case for deliberate practice. *Educational Psychologist*, *44*(4), 250–266.

Morphy, P., & Graham, S. (2012). Word processing programs and weaker writers/readers: A meta-analysis of research findings. *Reading and Writing*, *25*(3), 641–678.

National Center for Education Statistics (2012). *The Nation's Report Card: Writing 2011* (NCES 2012–470). Institute of Education Sciences, U.S. Department of Education, Washington D.C.

National Commission on Writing. (2004). *Writing: A ticket to work . . . or a ticket out*. New York, NY: The College Board. Retrieved from http://www.collegeboard.com/prod_downloads/writingcom/writing-ticket-to-work.pdf

Palermo, C. (2017). *A Framework for Deliberate Practice: Self-Regulated Strategy Development and an Automated Writing Evaluation Program.* (Doctoral dissertation). Retrieved from https://repository.lib.ncsu.edu/handle/1840.20/33737

Patthey-Chavez, G. G., Matsumura, L. C., & Valdes, R. (2004). Investigating the process approach to writing instruction in urban middle schools. *Journal of Adolescent & Adult Literacy*, *47*(6), 462–476.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53–76.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, *19*, 51–65.

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, *3*, 22–36.

Wilson, J. (2017a). Associated effects of automated essay evaluation software on growth in writing quality for students with and without disabilities. *Reading and Writing*, *30*(4), 691–718.

Wilson, J. (2017b). *Findings from analyses of Utah Compose and SAGE Data for academic year 2014-15.* Unpublished manuscript.

Wilson, J. (2017c). *Findings from analyses of Utah Compose and SAGE Data for academic year 2015-16.* Unpublished manuscript.

Wilson, J., & Andrada, G. N. (2016). Using automated feedback to improve writing quality: Opportunities and challenges. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 678–703). Hershey, PA: IGI Global.

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, *100*, 94–109.

Wilson, J., Olinghouse N. G., & Andrada, G. N. (2014). Does automated feedback improve writing quality? *Learning Disabilities: A Contemporary Journal*, *12*, 93–118.

Wilson, J., Olinghouse, N. G., McCoach, D. B., Andrada, G. N., & Santangelo, T. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing*, *27*, 11–23.