The Engine Driving Automated Essay Scoring

PEG®

Measurement Incorporated® (MI®) has been at the forefront of scoring student writing since the early 1980s. MI pioneered many of the complex processes involved in cost-effectively hand-scoring student essays on a large scale — scoring writing assessments for numerous US state departments of education, including Texas, Ohio, Michigan, Florida, and New Jersey. By the late 1990s, MI's expertise in hand-scoring had firmly established the company as the industry's premier writing assessment company.

By early 2000, MI had established a relationship with Duke Alumni Dr. Ellis Batten Page. Page is regarded as the "father of automated essay scoring" for his pioneering work in the early 1960s. He was the first to explore, document, and validate the computer-based assessment of written prose. His software was entering a new era as advances in microcomputer technology and the emergence of the World Wide Web were making automated essay scoring a practical possibility. In 2003, MI acquired Project Essay Grade® (PEG®) from Dr. Page and his associates. Thirteen years later, MI has re-engineered, enhanced, and extended the PEG system using the latest techniques and technologies in the field of computational linguistics, machine learning, and natural language processing.

With improvements in PEG and general advances in the reliability of machine scoring, automated essay scoring (AES) has become a valuable, and in some cases, essential, tool in a variety of contexts. MI's AES engine, PEG, is currently used in summative and formative assessments. It is being used in pilot and field tests for the Smarter Balanced Assessment Consortium (SBAC), which represents 31 states in the US. PEG's successful performance in pilot and field tests led to PEG being contracted in 2015 and 2016 to score hundreds of thousands of students' written responses, increasing the number of US states using PEG. PEG remains at the forefront of national assessment developments with its established track record in scoring essays for qualitative characteristics such as organization, support, word choice, and mechanics.

Latest Research

The PEG scoring platform utilizes new algorithms developed in-house that provably optimize the industry standard human-machine agreement metric known as quadratic weighted kappa (QWK). Human-machine agreement is the frequency of success of an automated scoring engine to produce the same, or comparable, scoring results as a human counterpart. Improving this metric of agreement tends to result in higher accuracy of the scoring application.

Automated scoring engines are typically trained using human-scored writing samples. We can then measure the accuracy of the AES engine by using it to score additional human-scored writing samples that the AES has never encountered before. By comparing these scores to the human scores, we can assess the agreement between the two. We have made several theoretical

advances that (to our knowledge) are new to the field and have led to a deeper understanding of QWK and its optimization in automated scoring.

The growing body of research and the increasing demand for large scale production scoring demonstrate the viability of AES scoring in general and MI's leadership in the automated scoring industry.

Summative Assessments

PEG's use in summative scoring has increased prominently in the last seven years.

Since 2009, the Utah State Board of Education has repeatedly used PEG successfully as the scoring method on the statewide summative Direct Writing Assessment in Grades 5 and 8. In the 2009-2014 timeframe, PEG scored 344,000 student responses on Utah's six trait rubric. In 2013 PEG was used as the second reader on the Connecticut SBAC Aligned Practice Assessment (APA), providing scores for 90,000 student responses on Connecticut's three trait rubric.

PEG's production scoring success in Utah and Connecticut as well as PEG's strong performance in the Smarter Balanced Field Test led to a request for a service allowing summative scores to be submitted to the PEG engine via automated transmission. This service request, based on client demand, led to the design of a streaming scoring service.

This service allows clients and partners to submit and receive scores for hundreds of thousands of responses per week. In 2015 PEG's Streaming Scoring web service went live. Throughout 2015 & 2016, and thus far in 2017, PEG Streaming Scoring has produced summative scores for well over fifteen million student responses from many partners and states in the US including California, Michigan, South Dakota, Vermont, and Wisconsin.

PEG's performance has been subject of considerable research of industry standards and among automated scoring competitors.

In Spring 2013, PEG was selected as one of the AI engines to be deployed by the Smarter Balanced Assessment Consortium to provide automated scoring of items on the pilot and field tests of its next generation assessments. PEG scored 213,000 essay and short answer (ELA and Math) responses for the pilot test in Fall 2013, and scored approximately 2.5 million responses for the field test in Fall 2014. Although many vendors scored subsets of the items, PEG was the only engine to score all tested items. PEG's results ultimately equaled or exceeded all engines tested. The PEG engine is listed as "Vendor 3" in the Smarter Balanced report: http://www.smarterapp.org/documents/FieldTest_AutomatedScoringResearchStudies.pdf)

MI anticipates PEG's involvement with upcoming Smarter Balanced field testing and research in the near future.

In 2012, the Hewlett Foundation sponsored global competitions in automated scoring – the Automated Student Assessment Prize (ASAP), Phases 1 and 2. These competitions were the first of their kind and were intended to independently evaluate state-of-the-art essay and short answer scoring. In both phases, PEG outperformed the competitors by achieving the highest level of agreement with respect to the human scores (Shermis & Hamner, 2013; Morgan, Shermis, Van Deventer, & Vander

Ark, 2013). In addition to the ASAP results, there is a wealth of independent research that examines the validity and reliability of automated scoring, particularly as it relates to summative assessment, including a large body of work conducted by Dr. Page himself over a span of nearly 40 years.

Formative Assessments

PEG also drives automated essay scoring for formative writing practice websites, and has been used to provide tens of millions of scores to students in formative writing assessments, with over six million essays scored in the last year alone. In addition to providing real-time scores, PEG also adds value when used in a formative context by providing response-specific feedback to students on the grammar and spelling errors found in their essays, as well as offering targeted instructional feedback on how to improve their writing skills. PEG is currently in use by North Carolina's NC Write, Educational Records Bureau's Writing Practice Program (WPP), Utah State Board of Education's Utah Compose, and MI's own PEG Writing and PEG Writing Scholar. PEG has also been used by previous writing programs including Connecticut State Department of Education's CBAS Write, Measurement Planet's Writing Planet, and Learning Express's Learning Express Advantage.

A study conducted at University of Connecticut's Neag School of Education has shown that students used PEG's automated scoring and feedback to increase their essay scores with repeated revisions of an essay, with the highest growth shown in the first few revisions (Wilson, Olinghouse, & Andrada). In related research, Wilson and Andrada used PEG's automated scoring and feedback to more accurately identify struggling writers, in comparison to a static first-draft assessment (Wilson & Andrada, 2013). Two-thirds of the students initially identified as at-risk, were able to move out of the at-risk classification given five or more revisions with feedback. These results point to the ability of PEG to not only assess writers in a typical summative assessment, but also to be used as an assessment and intervention tool in the context of a formative system.

Recent research at the University of Delaware (Wilson & Czik, 2016) explored the role of feedback in the improvement of both writing motivation and quality and discovered an increase in writing persistence in groups with feedback from both human teachers and PEG when compared to groups with feedback from human teachers alone. Teachers in the combined group also reported considerable time-savings in creating student feedback due to the contribution of PEG's automated feedback.

Morgan, J., Shermis, M. D., Van Deventer, L., & Vander Ark, T. (2013) Automated Student Assessment Prize: Phase 1 & Phase 2. Retrieved from http://gettingsmart.com/wp- content/uploads/2013/02/ASAP-Case-Study-FINAL.pdf

Shermis, M. D., & Hamner, B. (2013). Contrasting State-of-the-Art Automated Scoring of Essays. Handbook of Automated Essay Evaluation: Current Applications and New Directions, 313.

Wilson, J., & Andrada, G. (2013). Examining Patterns of Writing Performance of Struggling Writers on a Statewide Classroom Benchmark Writing Assessment: The Utility of Dynamic Assessment. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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