

A Comparison of Automated Essay Graders

How SAGrader distinguishes itself from other automated essay grading programs.



Value

SAGRADER™ IDEA WORKS	MY ACCESS!™ VANTAGE	WRITETOLEARN™ PEARSON	CRITERION™ ETS
<p>SAGrader™ is an online learning environment that maximizes students' engagement with course material through content-based writing assignments.</p> <p>By encouraging multiple drafts and providing directive, specific feedback, SAGrader helps students develop a strong understanding of course-specific concepts and the ability to communicate clearly and effectively.</p>	<p>With MY Access!™ students are motivated to write more and attain higher scores on statewide writing assessments. By using MY Access! in the classroom, teachers provide students with the practice they need to improve their writing skills.</p> <p>The program's IntelliMetric™ scoring engine grades students' essays instantly and provides targeted feedback, freeing teachers from grading thousands of papers by hand and giving them more time to conduct differentiated instruction and curriculum planning.</p>	<p>WriteToLearn™ is a complete online tool for building writing skills and developing reading comprehension.</p> <p>Using WriteToLearn, students develop a skill that good readers naturally possess - the ability to summarize what they read. They practice essay writing and summarizing and build both writing and reading comprehension skills across the curriculum.</p>	<p>The Criterion™ service is a web-based application that evaluates a student's writing skills and, within seconds, provides score reporting and diagnostic feedback to both writing instructors and students.</p> <p>Students draft and submit essays and receive immediate feedback in the form of a holistic score and diagnostic annotations within each essay that guides instruction.</p>

Scoring Model

SAGRADER IDEA WORKS	MY ACCESS! VANTAGE	WRITETOLEARN PEARSON	CRITERION ETS
<p>SAGrader explicitly models the knowledge, organization and reasoning that should be present in a good answer.</p> <p>Using a combination of artificial intelligence, natural language understanding, computational linguistics and fuzzy logic strategies, essays are scored based on the presence or absence of key features.</p> <p>Scoring does not rely on statistical models or patterns, but directly compares students responses to the semantic relationships modeled in the knowledge base, much like an actual instructor grades.</p>	<p>Vantage uses an inductive statistical model to predict good and bad essays rather than directly assessing knowledge.</p> <p>It examines over 400 features of text then uses a minimum of 300 human-scored essays to estimate parameters for a statistical model predicting whether new essays are more like good essays or bad essays.</p> <p>This statistical model does not actually read or analyze the specific content of entries. The scoring depends on how similar the entry is to previously graded submissions.</p>	<p>WriteToLearn uses a system based on Pearson's implementation of Latent Semantic Analysis (LSA), an approach that infers semantic similarity of words and passages by analyzing large bodies of relevant text.</p> <p>The Reading Comprehension Component assesses the correlation between the summary content and the original reading passage. The Essay Component assesses the correlation between the essay's content and that of training essays previously scored by human readers.</p> <p>Scoring relies on inferred semantic meaning based on the similarity between new entries and graded entries.</p>	<p>Criterion is based on a technology called e-rater®. E-rater scoring uses NLP to identify the features of faculty-scored essays in its sample collection and store them, with weights, in a database. When e-rater evaluates a new essay, it compares its features to those in the database.</p> <p>Because the e-rater scoring engine is not doing any actual reading, the validity of its scoring depends on the scoring of the sample essays in the e-rater database.</p>

Feedback

SAGRADER IDEA WORKS	MY ACCESS! VANTAGE	WRITETOLEARN PEARSON	CRITERION ETS
<p>By assessing specific substantive knowledge, SAGrader can give students detailed, personalized feedback indicating the specific learning objectives they have met and those needing further work.</p> <p>SAGrader measures how well a student understands concepts and the logical relationships among those concepts, not just general writing competency.</p> <p>Students receive assistive feedback written in natural language, which tells them exactly how to improve future submissions.</p>	<p>Vantage scores student submissions on 5 dimensions based on hundreds of specific features used to estimate each dimension (focus and meaning, organization, content and development, language use and style, mechanics and conventions).</p> <p>Student feedback includes scores on each broad dimension, with generic suggestions for how to improve. Feedback is unable to give students specific comments about precise concepts or how to improve.</p> <p>Feedback provides help with broad writing competency but cannot address if students understand specific concepts, theories, or issues within the substantive domain.</p>	<p>Students receive feedback on how well they covered each reading section along with feedback on spelling, copying from the text, repetition and inclusion of unimportant information, and the length of their summary.</p> <p>WriteToLearn gives only general feedback in the form of a graphic bar chart, scaling from 1 to 6. No specific information is given about the type of information students identified.</p> <p>Students receive only broad suggestions (if any) for improving their submission. This feedback is unable to address whether students understand specific concepts, theories, or issues within the substantive domain.</p>	<p>Students receive diagnostic feedback, as well as a holistic evaluation on a 4 or 6-point scale.</p> <p>Holistic scoring compares a student's writing to thousands of essays written and evaluated by writing instructors. Trait level indicators are also available. The traits for which information is provided are Grammar, Usage and Mechanics, Style Organization & Development.</p> <p>Since Criterion employs statistical modeling, students are only given general feedback regarding writing competency, not feedback about specific concepts, theories, or issues within the substantive domain.</p>

Accuracy

SAGRADER IDEA WORKS	MY ACCESS! VANTAGE	WRITETOLEARN PEARSON	CRITERION ETS
<p>Performance meets or exceeds that of human scorers. Accuracy for SAGrader is measured by correct recognition of each specific item, which simultaneously validates the whole score given.</p> <p>SAGrader accuracy can be objectively measured against an explicit rubric. This allows for an incredibly precise level of assessment, and permits users to see exactly how SAGrader's scoring and feedback matches the rubric.</p> <p>Grading and feedback accuracy is continuously monitored and improved over time, by adding semantic relationships to the knowledge base.</p>	<p>An independent 2008 study found that Vantage's IntelliMetric™ had no statistically significant correlation to the holistic scores given by human graders. Furthermore, it was shown to have a statistically significant correlation to human graders on just one scoring dimension - Sentence Structure.¹ This indicates that Vantage has little ability to rate essays holistically or for any specific area except sentence structure.</p> <p>Also, adjacent points on the 6 point scale are considered within the range of accuracy. This is a notoriously lax criterion (for categories 2-5, 50% agreement would be expected just by chance).</p>	<p>Performance meets or exceeds that of human scorers. However, adjacent points on a 6 point scale are considered within the range of accuracy. This is a notoriously lax criterion (for categories 2-5, 50% agreement would be expected just by chance).</p> <p>The correlation with human graders has been shown to be equal to or higher than that between two independent human graders in dozens of studies with over 200 prompts of every type. But these validation studies test the program with the same set of papers used to train it, making it impossible to generalize the measurement model.</p>	<p>Performance meets or exceeds that of human scorers. However assessment often considers adjacent points on the 6 point scale to be correct responses. This is a notoriously lax criterion (for categories 2-5, 50% agreement would be expected just by chance). They also report domain scoring is less accurate for Criterion than holistic scoring.</p> <p>Worse yet, the standard for assessment is not whether the essay successfully addresses important substantive issues but whether the essay looks more like good or bad essays.</p>

¹ Wang, J., & Brown, M. S. (2008). Automated essay scoring versus human scoring: A correlational study. *Contemporary Issues in Technology and Teacher Education*, 8(4), 310-325.

Quality Control

SAGRADER IDEA WORKS	MY ACCESS! VANTAGE	WRITETOLEARN PEARSON	CRITERION ETS
<p>Yes.</p> <p>Students are able to flag scores in a challenge communication system. This signals the instructor to evaluate the student's essay for a student error or an inaccuracy in the knowledge base. An inaccuracy can be quickly fixed to improve the evaluation of all students.</p> <p>Instructors can also proactively comment on student submissions to applaud good work or suggest improvements.</p> <p>The program is flexible for teachers and can fit changing demands and best practices.</p>	<p>No.</p> <p>Vantage offers no quality control mechanism once students begin using the program. There is no provision for student-instructor- developer communication or for students to express concerns.</p> <p>The program is not flexible for teachers and will not change to fit changing demands or best practices.</p>	<p>No.</p> <p>Pearson offers no quality control mechanism once students begin using the program. There is no provision for student-instructor- developer communication or for students to express concerns.</p> <p>The program is not flexible for teachers and will not change to fit changing demands or best practices.</p>	<p>No.</p> <p>ETS offers no quality control mechanism once students begin using the program. There is no provision for student-instructor- developer communication or for students to express concerns.</p> <p>The program is not flexible for teachers and will not change to fit changing demands or best practices.</p>