Beyond Automated Essay Scoring: Forecasting and Improving Outcomes in Middle and High School Writing

Elijah Mayfield, David Adamson, Bronwyn Woods,
Shayne Miel, Stephanie Butler, and Jill Crivelli
Turnitin, 2020 Smallman St, Pittsburgh PA 15232
{emayfield, dadamson, bwoods, smiel, sbutler, jcrivelli}@turnitin.com

ABSTRACT: This paper presents an analysis of an automated essay scoring (AES) system in two studies of live classroom use. First, in a study of 99 students in Texas, we show that automated scores do predict future performance on standardized tests, and that in-system activity can be included in a predictive model to further improve accuracy. Following that, the results of a five-school study in Maryland demonstrate moderate evidence that automated essay scoring is correlated with school-level improvement in ELA outcomes.

Keywords: Automated essay scoring, intelligent tutoring systems, writing assessment, school implementation, quasi-experimental efficacy

1 INTRODUCTION

Student performance in writing is difficult to assess at large scales, and targeted instruction based on that assessment is even more challenging. Unlike in math or reading, turnaround time for even short written student work can take weeks, and large-scale assessment for schools or districts may not be available until the following school year. Scoring relies on instructors or trained scorers who can become tired or distracted over hours of scoring, leading to inconsistent results (Williamson et al., 2012). English Language Arts (ELA) teachers at these grade levels can also teach up to 6 classrooms, up to 200 students at a time, which combined with low salaries and minimal support leads to particularly high attrition, low job satisfaction, and poor student outcomes (Scherff & Hahs-Vaughn, 2008).

Automated essay scoring (AES) aims to solve some of these problems. For half a century, researchers have have worked to reduce the time burden of (Page, 1966). This goal remains largely consistent today. AES models are trained on a small set of essays scored by hand, and then score new essays with the reliability of an expert rater. A large body of work, particularly in the last decade, has demonstrated this reliability (Shermis & Burstein, 2013).

This paper investigates AES in classrooms over time, evaluating the relationship of AES use to outcomes in two authentic settings. The first section studies the use of AES for *predicting* outcomes of individual students in a Texas high school. The second section evaluates whether AES *improves* ELA outcomes in several Maryland middle schools. This second causal claim is a much more challenging bar for AES than scoring accuracy or forecasting ability.

2 BACKGROUND

AES has focused historically on replicating expert readers for large-scale scoring of thousands of essays, either for end-of-year standardized assessments or entrance exams

like the GRE or TOEFL (Attali & Burstein, 2004). This use preferences interpretable model features informed by psychometrics, often representing high-level characteristics of writing like coherence or lexical sophistication. The primary goal is defensibility of the underlying model, known as construct validity. This construct validity through feature choice has been emphasized over measuring the ability to provide actionable guidance to writers based on the scoring.

In the 1990s and early 2000s, classroom technology was released based on this approach, including ETS Criterion, Pearson WriteToLearn, and Vantage MyAccess. Classroom reviews of these products were mixed at best. While their use positively impacted student writing (Shermis et al., 2008), students felt negative about the experience (Scharber et al., 2008). The most widely-cited districtwide study on these tools (Grimes & Warschauer, 2010) described the work as "fallible" and gains in school outcomes were not demonstrated. Teachers using earlier tools stated that automated scoring must be paired with actionable next steps for writers (Riedel et al., 2006). Building on this, work in academic settings has used AES to provide formative writing instruction and feedback that students perceive as "informative, valuable, and enjoyable" (Roscoe et al., 2013) and which provides more efficient learning gains than practice alone (Crossley et al., 2013).

Alongside the emergence of that research, a newer generation of tools has refocused AES to prioritize feedback to students. These include TenMarks Writing, WriteLab, Grammarly, PEG Writing, and Turnitin Revision Assistant. AES feedback's impact on writing quality varies by product. For instance, PEG Writing has been shown to save teachers time and let them focus on higher-level writing skills, but not to improve writing quality (Wilson & Czik, 2016). Revision Assistant provides feedback that students rate as helpful, and encourages editing that improves quality across drafts (Woods et al., 2017). To date, there is little work discussing the longitudinal effect of AES on classroom instruction during the school year.

2.1 Turnitin Revision Assistant

This research focuses on two school districts and one AES technology used in both, Turnitin Revision Assistant, released in 2016. Prior work has shown that the AES used in this product reliably predicts student writing scores in line with the state-of-the-art (Shermis, 2014).

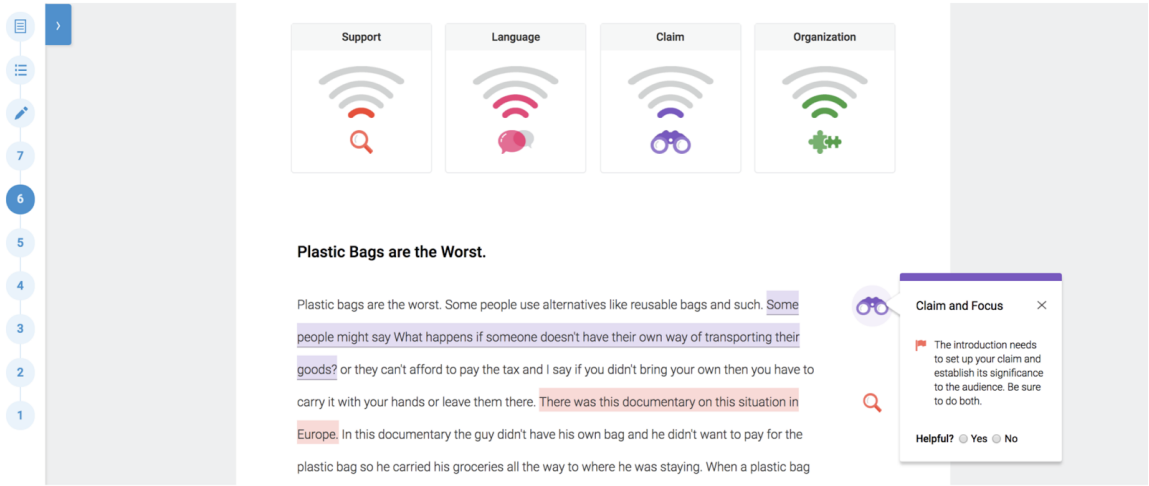

Figure 1: Automated essay scoring through Signal Checks in Revision Assistant.

In Revision Assistant, students request feedback from a "Signal Check". This provides automated scores on rubric traits in a visual format (Figure 1) and highlights up to four sentences within the text for in-line feedback. The full feedback algorithm is described in Woods et al (2017). Revision Assistant also contains "Spot Check" assessments, which remove real-time feedback, instead scoring essays for teacher review. This Spot Check environment matches summative settings like standardized testing and gives teachers insight into student skill transfer into settings where real-time support will not be available. Note that in this paper, a "draft" of student work corresponds to a Signal Check or a final submission; intermediate work between requests for feedback is not a separate draft.

Based on prior literature, AES feedback in Revision Assistant should have a positive impact on classrooms. Students who learn to think of writing as a process that includes iterative improvement demonstrate large gains in transferable skills (Dix, 2006; Tillema et al., 2011). Unfortunately, this process is difficult to learn and complex to teach, needing differentiated instruction across students and incorporating strategies that may vary across tasks (Hayes & Flower, 1980). Teachers tend to view this element of instruction as difficult and time-consuming, and rarely teach the revision process in depth (Graham & Harris, 2005).

3 FORECASTING STUDENT OUTCOMES

In Texas, student performance is evaluated on the State of Texas Assessments of Academic Readiness, or STAAR (Texas Education Agency, 2017). This test measures student progress against curricula aligned to Texas Essential Knowledge and Skills, or TEKS, standards. Students in grades 3-11 are assigned a Reading component, while writing is evaluated in the 4th, 7th, and 9th-11th grades. Writing Scores are broken out separately and are also combined with Reading scores into an overall ELA Score. This study evaluates the use of Revision Assistant to forecast student outcomes on the STAAR assessments on both the Writing Score and the combined ELA Score. We find that the Revision Assistant forecast compares favorably to, and effectively supplements the information provided by, the existing Fall benchmark currently used by the school.

3.1 Methods

Six English I (9th grade) classes from a large, urban school district, taught by four teachers, were selected to participate in the study. In January, teachers met and were trained on the AES system, including the difference between Signal Check and Spot Check assignments. A total of 111 students were enrolled in participating classes during the administration of a school-wide benchmark in fall 2017; of those, 99 participated in the study. Shortly after training, teachers administered an initial, timed Spot Check assignment. Teachers were then given access to Revision Assistant for three months, with a recommended pacing guide that included four writing prompts appropriate to the school curriculum and sequencing of English I. Adherence to this pacing guide was not mandated. A second Spot Check assessment was administered, no more than one month prior to the STAAR assessment. Spot Check assignments matched the genre of writing used in the STAAR assessment, though there were some differences. For instance, the writing was typed instead of handwritten, and was not subject to length constraints (Texas students are penalized for

exceeding a maximum length on standardized essay assessments). End-of-year STAAR testing, administered at the end of March 2017, was used for final performance evaluation.

We first evaluate the pre-existing benchmark and the results from Spot Check assessments as individual, linear predictors of STAAR performance. Next, we fit a multivariate linear regression using four variables and use that regression to predict overall STAAR ELA and STAAR Writing performance. In this model, the first three variables are direct student evaluations: the **pre-existing benchmark score**, student performance on the **initial Spot Check**, and student performance on the **second Spot Check**. The fourth is a measure from Signal Check assignments during the class curriculum - specifically, the **total count of Invalid Drafts** submitted by each student. An Invalid Draft is a draft that was not given a score, due to being off-topic or in bad-faith. Detection of such drafts is fully automated through machine learning. Invalid drafts can represent student "churn" - an inability to compose essays that meet assignment criteria - or student disengagement. Both are early warning signs that can be addressed through targeted instruction.

We evaluated other factors from formative assignments in a Signal Check setting, such as total number of drafts authored, growth (as measured by automated scoring) within-assignment, and increases in word count. In a forward stepwise regression, after including variables for student performance on Spot Check assignments, none of these other factors improved model fit or were significant in a *t*-test. They are not included in our results.

3.2 Results

The school district's existing fall benchmark is reasonably reliable at forecasting student performance on the STAAR ELA assessment as a whole (r = 0.63). However, it is only slightly predictive of student end-of-year STAAR Writing performance (r = 0.26). As this has historically been the only available benchmark, the school has had no access to actionable insights on student writing performance. The first Spot Check, by contrast, was moderately predictive both of STAAR Writing (r = 0.45) and overall STAAR ELA (r = 0.43) performance.

Table 1: Accuracy of models by correlation (r) and root mean squared error (RMSE)

	STAAR Writing (2-8 scale)		**STAAR ELA (0-68 scale)**	
	r	RMSE	r	RMSE
Existing Benchmark	0.26	0.97	0.63	6.63
Spot Check Assessment	0.45	0.90	0.43	7.70
4-Variable Forecast	0.58	0.82	0.74	5.42

In the multivariate model, the addition of AES variables significantly increases predictive accuracy over the benchmark alone, which combine to explain 55% of student performance (r^2) on STAAR ELA. The four variables are all significant ($p < 0.01$) in both forecasting models. This analysis is summarized in Table 1; scatter plots are presented in Figure 2. These results give evidence of the value of AES for forecasting end-of-year student performance.

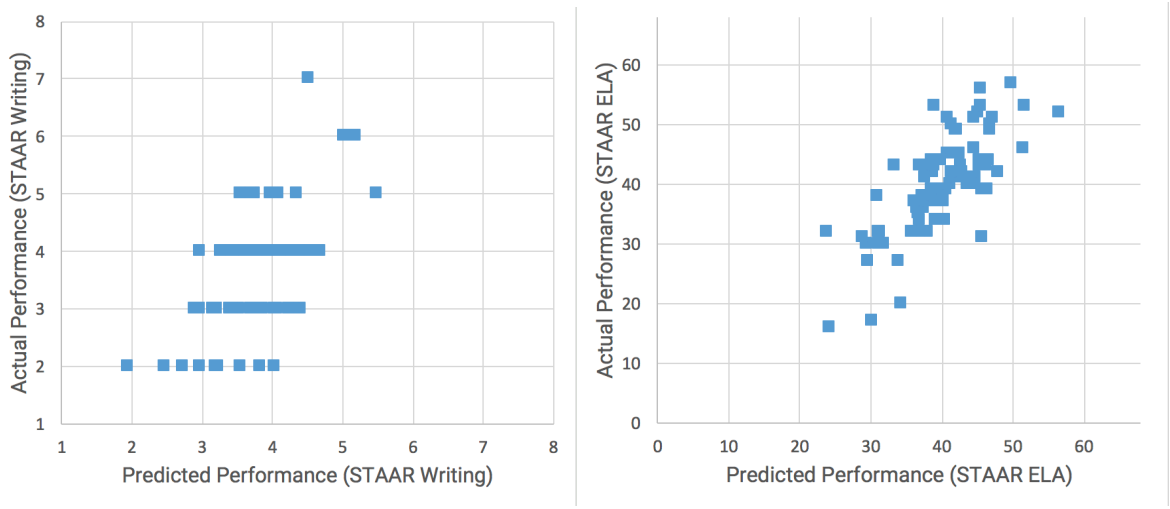

Figure 2: 4-Variable Forecast of scores on STAAR Writing (left) and STAAR ELA (right).

4 IMPROVING SCHOOL OUTCOMES

The next study evaluates a more challenging benchmark: whether use of AES within a standard ELA curriculum improves outcomes over time. To study this, we conducted a multi-site, quasi-experimental study of middle schools in a large, rural school district in Maryland.

4.1 Methods

Teachers in the school district were provided with unlimited access to Revision Assistant during the school year. In fall 2016, trainings were conducted on-site in large groups, and virtually in smaller groups. Staff provided resources to teachers that aligned content in Revision Assistant to school curricula. No specific pacing was mandated by the district. Five schools participated in a treatment condition using Revision Assistant in their curriculum, while two schools in the district did not participate.

Maryland is a consortium member of the Partnership for Assessment of Readiness for College and Careers, or PARCC, which authors end-of-year assessments based on the Common Core State Standards (Maryland Department of Education, 2017). School performance was measured using the PARCC end-of-year assessment for English Language Arts students in 8th-grade, the final year before entrance to high school. Students who exceed or meet expectations, the top two performance categories, are defined as passing.

Table 2: Usage statistics for five participating schools in the 2016-2017 school year.

School	Signal Checks	# Drafts / # Submissions	Mean Increase in Summed Score	Mean Increase in Word Count
1	2,011	14.1	5.5	636
2	3,187	9.9	2.7	788
3	596	5.6	1.9	194
4	6,155	11.3	3.0	413
5	4,733	11.0	2.1	294

Table 3: Change in 8th-grade ELA pass rates in treatment schools.

School	2016	2017	Change
1	35.8	49.1	+13.3%
2	41.1	49.0	+7.9%
3	39.0	45.8	+6.8%
4	30.7	35.5	+4.8%
5	23.7	22.2	-1.5%
Treatment Average (n=5)	34.1	40.3	+6.2%
Non-Treatment Average (n=2)	40.4	38.7	-1.7%
Maryland Average (n=352)	31.8	32.0	+0.2%

4.2 Results

Table 2 describes Revision Assistant usage overall within participating schools. Four schools show high activity while one shows lower levels of activity. In all schools, students composed many drafts prior to submitting their work, represented by the drafts per submission count. The summed score is summed over rubric traits, and so is on a scale of 4 to 16. Mean number of drafts per student submission, mean increase in summed score, and increase in word counts all broadly replicate the finding from Woods et al (2017) of students receiving automated feedback and subsequently improving their essays; however, any one of these measures in isolation is incomplete in capturing student growth or essay quality.

Table 3 presents performance of the five schools in the PARCC exam. Average increase among these schools was 6.2%. By contrast, passing rates declined in both non-treatment schools in the district, by 0.2% and 3.2%. Statewide, average change in pass rates from 2016 to 2017 was an increase of 0.2% (PARCC did not administer an exam prior to 2016, so no further historical data is comparable). The scatter plot in Figure 3 places these schools in the statewide context of all 352 middle schools.

To evaluate the significance of the high rate of improvement in treatment schools, we conducted a permutation test, randomly sampling subsets of five schools from the full population. This lets us evaluate the probability of five arbitrary schools showing similar growth by chance, though it does not account for any potential confounding factors driving both testing success and Revision Assistant usage. These sampled subsets of schools showed mean change greater than 0% in 51% of random samples, amounting to a coin flip. Permutation subsets of schools with mean growth over 6.2% were rarer. Subsets matched or exceeded the performance observed in Revision Assistant schools in 6% of simulations, indicating a 6% chance of these results being observed by chance.

5 DISCUSSION

Combined, these two studies present evidence of the effectiveness of AES in school settings. The first study demonstrates forecasting power of AES to predict student outcomes. The second study demonstrates moderate evidence for improved outcomes and student growth tied to the deployment of an AES product in classrooms during a school year.

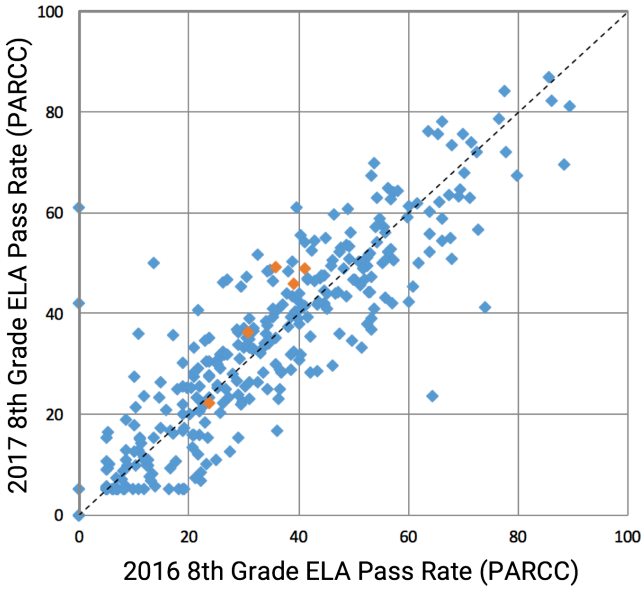

Figure 3: Performance growth of 5 treatment schools (orange) against all other MD schools (blue). The diagonal dashed line represents no year-over-year change.

Based on these results, we recommend two possible paths for schools using AES for forecasting purposes. For locations capable of administering and collecting a full benchmark replication of end-of-year assessments, that process continues to have value. When combined with formative AES activity, the overall predictive accuracy is high for student writing performance and very high for overall ELA performance.

However, the administration of a full benchmark assessment is time-consuming and distracting. For schools without the resources or time for these benchmarks, the results in this work suggest that a single, lightweight AES assessment is a moderately reliable indicator on its own, and adds minimal scoring overhead. In either case, these results are available early in the school year and provide time for targeted intervention. Moreover, the results of the Maryland study suggest a positive impact of AES in year-long classroom use.

5.1 Limitations and future work

Both studies have significant limitations. Neither study was subject to random assignment, relying on volunteers and self-selection. Furthermore, teachers were not subject to a rigorous pacing guide. The impact of AES may therefore be confounded with pre-existing differences, such as teacher readiness for adoption of educational technology, variations in funding of individual schools, or preparedness of building-level instructional coaching staff. Further research will require replication of these results with controlled assignment of students to conditions. The collection of student metadata will remove additional confounds and allow the evaluation of AES systems in light of recent work in fairness of machine learning systems (Leidner & Plachouras, 2017). Furthermore, future research could investigate *individual* student outcomes, an even more granular result not yet studied here.

REFERENCES

Attali, Y. & Burstein, J. (2006). Automated Essay Scoring With E-rater V. 2.0. *Journal of Technology, Learning, and Assessment (JTLA)*, *4*(3).

Crossley, S. A., et al (2013). Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. *International Conference on Artificial Intelligence in Education* (pp. 269-278). Springer, Berlin.

Dix, S. (2006). "What did I change and why did I do it?" Young writers' revision practices. *Literacy*, *40*(1), 3-10.

Graham, S., & Harris, K. (2005). *Writing Better: Effective Strategies for Teaching Students with Learning Difficulties*. Brookes Publishing Company. Baltimore, MD.

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).

Hayes, John & Flower, Linda. (1980). Identifying the organization of writing processes. Cognitive Processes in Writing (3).

Leidner, J. L., & Plachouras, V. (2017). Ethical by Design: Ethics Best Practices for Natural Language Processing. *European Association for Computational Linguistics 2017*, 8.

Maryland Department of Education. (2017). PARCC for Parents & Teachers. Retrieved from http://www.marylandpublicschools.org/programs/Pages/Testing/PARCC/index.aspx

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238-243.

Riedel, E., Dexter, S. L., Scharber, C., & Doering, A. (2006). Experimental evidence on the effectiveness of automated essay scoring in teacher education cases. *Journal of Educational Computing Research*, *35*(3), 267-287.

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2013). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition*, *34*, 39-59.

Scharber, C., Dexter, S., & Riedel, E. (2008). Students' experiences with an automated essay scorer. *The Journal of Technology, Learning and Assessment*, *7*(1).

Scherff, L., & Hahs-Vaughn, D. L. (2008). What we know about English language arts teachers. *English Education*, *40*(3), 174-200.

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.

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Shermis, M. D., Garvan, C. W., & Diao, Y. (2008). The Impact of Automated Essay Scoring on Writing Outcomes. *National Council of Measurement in Education.* New York, NY.

Texas Education Agency. (2017). STAAR Performance Standards. Retrieved from https://tea.texas.gov/student.assessment/staar/performance-standards/

Tillema, M., van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2011). Relating self reports of writing behaviour and online task execution using a temporal model. *Metacognition and Learning*, *6*(3), 229-253.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, *31*(1), 2-13.

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.

Woods, B., Adamson, D., Miel, S., & Mayfield, E. (2017). Formative Essay Feedback Using Predictive Scoring Models. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining* (pp. 2071-2080). ACM.