Districtwide Implementations Outperform Isolated Use of Automated Feedback in High School Writing

Elijah Mayfield, Turnitin, elijah@turnitin.com
Stephanie Butler, Turnitin, sbutler@turnitin.com

Abstract: This paper describes a large-scale evaluation of automated writing evaluation in classroom, non-high-stakes settings. Thirty-three high schools in California se of an educational product, *Turnitin Revision Assistant*, during the 2016-2017 school year. We demonstrate moderate evidence of growth in student outcomes based on this usage in general, exceeding rates of improvement statewide. We empirically demonstrate that broader adoption across buildings within a school district is correlated with stronger outcomes, and discuss implementation steps that can support those broad adoptions. Finally, we replicate this finding in a new context, with a full district adoption case study in Georgia, comprising ten schools.

Introduction

With the American Recovery and Reinvestment Act of 2009, the Department of Education provided $4.35 billion in additional grant funding to state and local school systems through the *Race to the Top* program (Duncan, 2015). Much of this supplemental funding went to educational technology, and led to further public and private investment in the industry. Combined, for-profit and non-profit financing in ed-tech peaked at over $4 billion in 2015 (Watters, 2016), mostly from a few large funders, like the Bill & Melinda Gates Foundation.

As investment increases, a consensus opinion has formed that not all educational technology has positive impacts on the classroom. Technology use in classrooms has been divided by socioeconomic status and regional access since the early years of its introduction into schools (Warschauer et al., 2004). These dividing lines are often tied to race, class, and other non-academic factors (Charity-Hudley & Mallinson, 2015). As a result, demand has surged for proof of broad, equitable efficacy and impact in authentic settings when evaluating technology interventions. We know that today, "although Ed Tech developers value research to inform their products, they are not conducting rigorous research" (Hulleman et al., 2017). This ties into a broader understanding, forming in parallel, that modern algorithmic products are learning to automated behaviors based on biased or unfair training data (Caliskan et al., 2017).

Part of the challenge in evaluating education research is that technology products behave differently in implementation at scale than they do in lab settings. Schoolwide use of an intervention can introduce difficulties that are not present in controlled environments. Products designed for use in one context may sometimes be used in entirely different ways by communities with established behaviors and practices (Ogan et al., 2012). Additionally, products well-grounded in learning science theory can show large improvements in student outcomes in controlled settings (Aleven & Koedinger, 2002), while producing weak or even null results in broad adoptions (Cabalo et al., 2007). These results produce critical reporting and skepticism of educational technology products (Gabriel & Richtel, 2011).

This paper evaluates whether the learning outcomes of one intervention transfer from small-scale trials to large-scale implementation. We study *Turnitin Revision Assistant (RA)*, a machine learning-powered automated writing evaluation product, primarily designed for formative classroom use in American middle and high schools. We seek not only to gather positive evidence of efficacy, but to understand the conditions under which implementations of technology produce improved student outcomes.

Intervention Description

RA supports students in the writing process with feedback based on automated essay scoring (AES) algorithms, a widely adopted machine learning application that affects millions of students each year through tests like the GRE and GMAT (Ramineni et al., 2012). The product has been shown to reproduce expert scoring of student essays reliably (Shermis, 2014), and generate feedback that improves student work (Woods et al., 2017). Early results have shown improved outcomes for participating schools (Mayfield et al., 2018).

RA emphasizes the importance of the writing process by reframing essay authorship as an on-going activity. The design of the system utilizes AES to embed an intensive revision process into student interactions with the system. As students request automated scoring, feedback is also provided; *RA* highlights two relatively strong sentences and two relatively weak sentences (Woods et al. 2017). Instructional content appears alongside those sentences that helps students understands where they are excelling in their writing and where they should focus their revision efforts. Comments encourage students to take small, targeted steps toward iteratively

improving their writing. This design and pedagogical constraint is meant to provide students with the opportunity and the desire to engage in writing strategies around constant refinement and iteration.

By creating an environment that directly connects student writing to feedback that encourages rework, it becomes clear to the student that good writing is the product of multiple drafts. The instantaneous nature of the feedback further aids students by creating an environment where revision can easily take place. Feedback cycles which could be days or weeks long are shorted to near-instantaneous feedback. This makes it significantly more motivating for students to revise and improve their work. The visual, game-like appeal of Wifi signals creates an atmosphere that encourages students to work and improve, without the emphasis on assessment and testing that has been present in previous, summative AES systems.

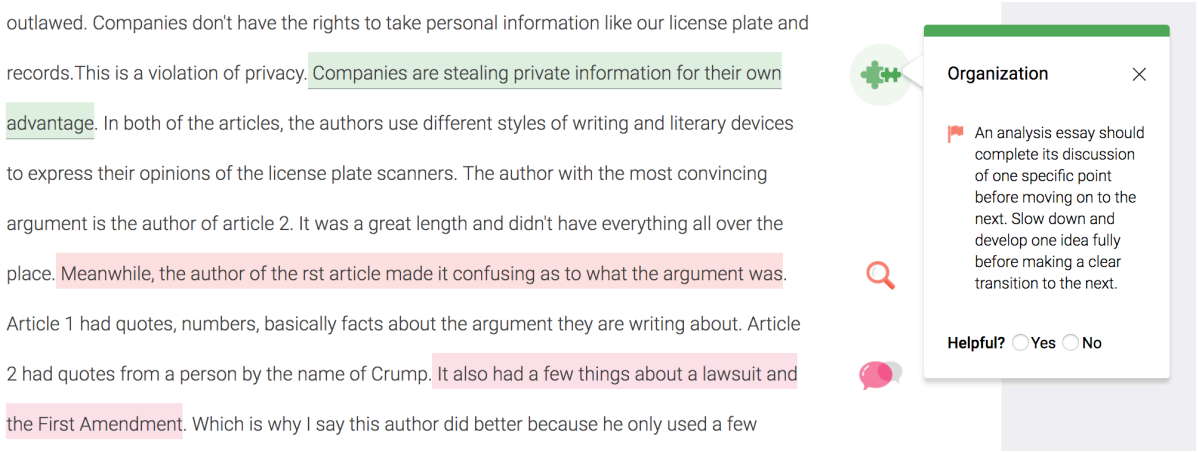

Figure 1. In-line feedback supplements automated, trait-based scoring in *Revision Assistant*.

Practitioner Involvement in Curriculum and Content

Content within *RA* is organized into prompts, separated by genres, in a prompt library. Prompts are curated, edited, and contextualized with supporting materials like exemplar essays by a team of curriculum specialists with prior teaching experience. Teachers and schools cannot directly import new content into the product or assign new open-ended writing tasks; instead, content is vetted and expanded in partnership between practitioners and product developers to maintain validity of the underlying algorithms (West-Smith et al., 2018).

The original source of content varies. In some cases, school districts propose initial drafts of content directly from pre-existing assignments; at other times, partnerships are developed with other intervention developers or curriculum companies to support licensed content like the Advanced Placement (AP) program. Finally, some content and most supporting materials are written entirely by the in-house curriculum specialists; when supporting materials are authored by teachers, they are compensated for their work and attributed with a byline on the resulting materials. Each writing prompt in the *RA* prompt library is associated with a fixed genre, rubric, and set of appropriate grade levels for assigning the task. In order for automated scoring to be reliable and automated feedback to be aligned to the wording of the rubric, creation of new prompts is not supported; this is in contrast to other AES products, which often have an open-ended scoring mode that defaults to more generic and surface-level feedback and scoring (Attali et al., 2010).

Implementation and Training

No single-teacher or individual classroom access to *RA* is available; the minimum registration is at the level of individual schools, and district-level implementations are common. As a result, prior to teachers using *RA*, school or district administration have typically developed an implementation plan in coordination with Turnitin. Superintendents, principals, or other administrators agree to have prompts aligned to their local curriculum units, often as a partnership instructional coaches and specialists within Turnitin. This approach creates a consistent sequencing of content for students across classrooms, and avoids duplication of the same assignment to a student by multiple teachers with access to *RA* over time.

Most schools that purchase *RA* schedule some form of training or professional development prior to wide use. The two most common formats are a 60-minute virtual training and a 3-hour on-site training. In the former case, the training is procedural – users learn the premise of the product and receive instructions on how to browse the prompt library, create assignments, view student work, and download reports of student progress. The longer, three-hour sessions are presented by a Turnitin-employed specialist, visiting a school and

additionally providing broader context on student engagement in the writing process, strategies for revision, and classroom lesson planning guidance for use of the technology. In larger adoptions – particularly districtwide adoptions – teachers also receive an introduction on how the product is sequenced with their school's curriculum and how they should collaborate across classrooms during the school year.

Teachers use *RA* in multiple ways: as a timed, in-class assessment; an untimed in-class activity with a take-home component; or as an entirely take-home, multi-day activity. In most cases, students have at least one full class period to use *RA* on laptops or in computer labs. They are then permitted to continue their work at home over the next few nights. When students are using Revision Assistant, they are always logged into an online interface. At any time after their first draft receives Signal Checks, students may press the Turn In button in the Revision Assistant interface, which lets their teacher know that the assignment has been submitted for review and, if it is being graded, for grading. Revision Assistant does not accept file uploads, though some students choose to write drafts in Google Docs or Microsoft Word and then paste the contents into the Revision Assistant interface only for feedback. Teachers may always review the roster for an assignment, where they can see each student's progress, draft history, and final submissions.

Methods

This study evaluates the performance of all California schools that were customers of *RA* during the 2016-2017 school year. Every school with 11th-grade students was evaluated, totaling 33 schools in 16 school districts. The end-of-year, high-stakes assessment for these schools is the California Assessment of Student Performance and Progress (CASPP). We track six student outcome measures from CASPP in our results – overall passing rates on the English Language Arts (ELA) portion of the test; percentage of students achieving the top scoring category on that test, *Exceeds Expectations*; and passing rates of students on each of the four broken-out ELA subskills: Reading, Writing, Speaking, and Research. The CASPP test was overhauled in 2015 to account for new standards in line with the Common Core; California is a member of the Smarter Balanced Assessment Consortium. As a result of this overhaul, long-term trends in performance are not comparable.

Prior to analysis of performance within *RA* schools, we evaluated differences between participating schools and the statewide average of all schools. Table 1 indicates that Revision Assistant schools in the treatment school year tended to enroll more students, with better pre-existing results both in CASSP passing rates and overall high school graduation. This self-selection is important for any interpretation of results in this paper, which are quasi-experimental by nature and do not represent a randomized trial across high schools.

Table 1: Differences between treatment schools and statewide schools prior to the 2016-2017 school years.

	RA High Schools	Other CA High Schools	Δ
Incoming ELA Pass %	66.1%	50.6%	+15.5%
11th-Grade Enrollment	527	251	276
Graduation Rate	95.1%	81.7%	+13.4%

We also measured implementation size and total usage of Revision Assistant. Four implementations were multi-school and districtwide, accounting for 15 schools. Eighteen schools implemented Revision Assistant as a partial adoption; nine were implemented as single schools, while nine more were partial adoptions of two or three schools within a school district. Usage across all participating schools was moderate to high; a total of 170,651 essay drafts were composed in California schools during the 2016-2017 school year, with large variance in per-school usage ($\mu = 5171$, $\sigma = 6676$). Beyond the guidance and partnerships described in the implementation design sections earlier in this paper, Turnitin did not specify a mandatory curriculum sequence for teachers to use *RA*, nor did Turnitin mandate any particular pacing or usage levels from teachers.

Because this study was performed at such large scales, and in fully authentic environments where student privacy requirements restricted access to individual-level data, this study only evaluates outcomes in aggregate at the level of school statistics; we do not analyze outcomes at the individual student level.

Results

Based on the sample size collected, improvements of 0.6σ or greater were statistically significant. When divided by implementation size, partial adoptions of Revision Assistant (either at single schools or partial district adoptions) showed no significant changes in school outcomes compared to statewide averages. By contrast, districtwide adoptions of *RA* showed significant growth in several outcome metrics: the number of students receiving the top score of *Exceeds Expectations*; passing rates in reading; and passing rates in speaking. Overall

passing rates improved in districtwide adoptions by 3.3%, compared to 0.7% growth in passing rates statewide; subscore performance ranged from a 0.8% growth in Research (no different from statewide average) to a 9.3% growth in passing rates on the Reading subscore (compared to 4.4% growth statewide). These results are summarized in Figure 2.

Figure 2. CA school growth in student outcomes from 2016 to 2017 (with effect sizes in σ)

Among *RA* schools, most additional variables were not significant predictors of year-over-year improvement. Neither pre-existing school performance indicators (like incoming 2016 graduation rate and incoming 2016 passing rates) nor quantity-based metrics of *RA* usage (such as total number of Signal Checks by all students) were significant predictors of growth in school performance. Among variables tested, only total school enrollment was a significant predictor of growth in performance among *RA* schools (r = -0.34, $p < 0.05$). This suggests that smaller high schools may have achieved higher rates of improvement compared to larger high schools. However, because of the number of hypotheses tested, this should be considered only a mildly predictive source of evidence and guidance for future research.

Based on these results, further details about the four districtwide implementations of *RA* are given in Table 2. Those districtwide adoptions took place in relatively large school districts with passing rates and graduation rates above the statewide average; three of the four districts had free and reduced lunch participation rates below the statewide average, while one was above the statewide average. All four districts showed improved passing rates overall; however, in three of four districts, the largest shift was an increase in the number of students performing at the top level, *Exceeds Expectations*, greater than overall passing rates.

Table 2: Statistics and performance breakdown for the four districtwide *RA* implementations.

	Students	2016 Pass %	Grad %	FRL %	Exceeded	Met	Nearly Met	Not Met
1	20-25,000	63%	93%	45%	29.7 (+4.7)	34.5 (-3.5)	22.0 (-1.0)	13.8 (-0.2)
2	10-15,000	66%	93%	46%	38.6 (+5.6)	30.9 (-2.1)	18.2 (-2.8)	12.3 (-1.7)
3	10-15,000	60%	96%	69%	25.0 (+1.0)	38.3 (+2.3)	22.2 (-0.8)	14.6 (-1.4)
4	30-35,000	78%	93%	17%	55.3 (+6.3)	25.5 (-3.5)	10.6 (-3.4)	8.7 (-0.3)
CA statewide		**59%**	**83%**	**59%**	**27.7 (+1.7)**	**32.0 (-1.0)**	**21.3 (-0.7)**	**18.9 (-0.1)**

Replication

The outcomes of districtwide implementations above replicate the positive findings of an implementation across five schools in Maryland (Mayfield et al., 2018). To gather additional evidence, data was gathered on an exemplar district-wide implementation of *Revision Assistant* in Georgia. Experimental methods were identical to those described above; only one district in Georgia used *RA* in the 2016-2017 school year. The district serves over 30,000 students and has 10 high schools, all of which were given access to *Revision Assistant.* The school district represents a medium-sized city with a wide range of socioeconomic status and pre-existing performance levels. High school students are assessed using the Georgia Milestones Assessment System in 9th and 11th grade English Language Arts. *RA* was embedded in writing curriculum across the district along with extensive administrator support, professional development and training for teachers.

The findings from districtwide implementations in California were replicated. Across the schools, student passing rates improved by 0.6σ in 9th grade and 0.4σ in 11th grade, compared to statewide performance. These results may be tempered by ceiling effects; the two highest-performing schools, as shown in Table 3, had pre-existing passing rates above 90%, leaving little room for growth. Within the same district, passing rates at other high schools were as low as 5% in 2016. Further confirming findings from California, raw *RA* usage analytics were not predictive of outcome within schools; in fact, there was a slight negative correlation between quantity of usage and growth in performance (r = −0.29 and −0.24 for 9th and 11th grade). This may be confounded with the ceiling effect for high-performing schools, where usage of *RA* was very high.

Table 3: Year-over-year change in ELA performance after full-year adoption of *RA* in the replication district.

	9th 2016	9th 2017	Growth	11th 2016	11th 2017	Growth
1	22.0	33.7	+11.6	13.8	32.2	+18.4
2	29.4	35.7	+6.2	27.6	44.9	+17.4
3	22.3	21.2	-1.1	15.8	32.1	+16.3
4	43.0	55.6	+12.6	36.6	45.9	+9.3
5	14.8	23.8	+9.0	7.5	15.5	+8.0
6	10.2	26.3	+16.1	16.2	22.4	+6.2
7	2.1	14.3	+12.2	5.0	9.7	+4.7
8	7.3	19.0	+11.7	6.1	8.0	+1.9
9	75.6	88.1	+12.5	86.9	88.3	+1.4
10	92.7	95.5	+2.8	94.8	90.5	-4.3
District (Average)	**25.0**	**34.3**	**+9.3**	**22.1**	**31.1**	**+9.0**
Georgia (Average)	**70**	**74**	**+4**	**65**	**70**	**+5**

Analysis

The results above unambiguously demonstrate, through large-scale quasi-experimental evaluation, that districtwide adoptions of *RA* produce significant improvements in student outcomes at the school level. The study does not provide evidence that partial adoptions or isolated use of *RA* produces long-term outcomes. This gap may explain results that have been highlighted in previous literature on AES. Specifically, while prior work has shown that ongoing practice of writing with AES improves student outcomes in experimental settings (Roscoe & McNamara, 2013), classroom implementations over school years have more often found positive impacts on teachers but no improvement in student outcomes (Wilson & Czik, 2016). Studies of districtwide implementations are rare, but have found the tools to be useful, but "fallible" (Grimes & Warschauer, 2010) with serious complications in deployment and real-world use. Our results suggest differential performance based on breadth of implementation, but does not yet establish a causal factor for this gap. Below, we propose some potential explanatory mechanisms and encourage further research.

A straightforward explanation is that classroom implementation of AES requires curriculum alignment, professional development, and district administrative support as a necessary causal component for efficacious ed-tech interventions. In this explanation, the additional work performed for district-level adoptions gives teachers the resources needed to make good use of technology in student interactions. These factors are invisible in quantity-based usage statistics, which did not predict student outcomes in this study. If this is a primary causal factor, it may land in a blind spot of typical quantitative outcome measures, a side effect of the "data deluge" in modern education (for more on the implications of this, see de Freitas & Dixon-Román, 2017).

An alternate explanation is economic. *RA* is sold to school districts and districtwide implementations require a larger allocation of school funding than isolated use. An explanatory hypothesis based on this theory would aver that the purchase of educational technology acts as a forcing function for improved collaboration and alignment among teachers and administrators, within and across schools. This improved teamwork would lead to better student outcomes, based on district commitment. Verifying this would need to explain any incongruence between this hypothesis and previous research showing that financial incentives, like teacher pay-for-performance, had minimal explanatory power in predicting student outcomes (Springer et al., 2011).

In light of this complex interplay between engagement and outcomes, further research will benefit from a mixed-methods approach that ties quantitative metrics to a richer qualitative understanding of implementations. Translational research will need to verify that experimental results are positioned in school district settings in a way that will reproduce those results. Researchers in the learning sciences may benefit from a mixed-methods approach that gives considerable weight to lived experience of practitioners. School administrators, in turn, will benefit from more nuanced reading of research results that lean less on individual metrics, especially of raw engagement, and more on the broad adoption of an intervention across teachers and schools. of an implementation. For all stakeholders, more work is needed to build a refined understanding of the behaviors within an algorithmic product that lead to transferable outcomes for students and teachers.

Next Steps

The schools represented in the *RA* population are not a representative sample of students in California or Georgia; instead, they are self-selected. The "gold standard" next step would be to replicate this analysis in a

fully experimental setting with schools subject to random assignment into conditions. However, fully randomized experiments at the scale described in this paper is beyond the scope of most learning sciences research. For instance, researchers may consider matched population techniques such as propensity score matching in future evaluations, to build a pseudo-randomized control population without requiring a full trial. Additionally, traditional learning sciences research into causal mechanisms or explanations at the level of individual students (either cognitive or sociocultural) will continue to require the smaller-scale, targeted experiments. Large-scale evaluations do not replace such endeavors, though they build a foundation for the continued transfer of research out of the laboratory and into the practitioner's classroom.

References

Aleven, V. A., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive science*, 26(2), 147-179.

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *Journal of Technology, Learning and Assessment*, 10(3).

Cabalo, J. V., Ma, B., & Jaciw, A. (2007). Comparative Effectiveness of Carnegie Learning's" Cognitive Tutor Bridge to Algebra" Curriculum: A Report of a Randomized Experiment in the Maui School District. Research Report. Empirical Education Inc.

Caliskan, A., Bryson, J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186.

Charity-Hudley, A., and Mallinson, C. (2015). *Understanding English language variation in US schools.* Teachers College Press.

de Freitas, E., & Dixon-Román, E. (2017). The computational turn in education research: Critical and creative perspectives on the digital data deluge. *Research in Education,* 98(1), 3-13.

Duncan, A. (2015). Fundamental Change: Innovation in America's schools under Race To The Top. US Department of Education.

Gabriel, T., and Richtel, M. (2011). Inflating the software report card. The New York Times, A1.

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).

Hulleman, C., Burke, R., May, M., Charania, M., and Daniel, D. (2017). Merit or Marketing?: Evidence and quality of efficacy research in educational technology companies. White paper produced for the *University of Virginia EdTech Academic Efficacy Symposium*.

Mayfield, E., Adamson, D., Miel, S., Woods, B., Butler, S., & Crivelli, J. (2018). Beyond Automated Essay Scoring: Forecasting and Improving Outcomes in Middle and High School Writing. *International Conference on Learning Analytics and Knowledge.*

Ogan, A., Walker, E., Baker, R., Rebolledo Mendez, G., Jimenez Castro, M., Laurentino, T., et al. (2012). Collaboration in cognitive tutor use in Latin America: Field study and design recommendations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Ramineni, C., Trapani, C., Williamson, D., Davey, T., & Bridgeman, B. (2012). Evaluation of the e-rater® Scoring Engine for the GRE® Issue and Argument Prompts. ETS Research Report Series, 2012(1).

Roscoe, R., & McNamara, D. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, 105(4).

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

Springer, M., Ballou, D., Hamilton, L., Le, V., Lockwood, J., McCaffrey, D., et al. (2011). Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT). *Society for Research on Educational Effectiveness.*

Warschauer, M., Knobel, M., and Stone, L. (2004). Technology and equity in schooling: Deconstructing the digital divide. *Educational Policy* 18:4. 562–588.

Watters, A. (2016). The Business of Ed-Tech 2016. Accessed on January 29, 2018 from http://2016trends.hackeducation.com/business.html.

West-Smith, P., Butler, S., & Mayfield, E. (2018). Trustworthy Automated Essay Scoring without Explicit Construct Validity. *AAAI Spring Symposium on AI & Society: Ethics, Safety, and Trustworthiness in Intelligent Agents.*

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.

Woods, B., Adamson, D., Miel, S., & Mayfield, E. (2017). Formative Essay Feedback using Predictive Scoring Models. *International Conference on Knowledge Discovery and Data Mining*.