Journal of Writing Assessment

Ernst, D. (2025). The effects of automated writing evaluation technology on improving student writing. *Journal of Writing Assessment*, *18*(2). **DOI**: 10.5070/W4.jwa.284

The Effects of Automated Writing Evaluation Technology on Improving Student Writing

Daniel Ernst, Texas Woman's University, US, dersnt@twu.edu

Abstract: Advances in automated writing evaluation (AWE) technology have shifted the aims of the tools from summative and holistic scoring and ranking essays to providing formative and analytical feedback to users for improving writing. This study uses a quasi-experimental design to test the ability of one AWE program to improve college student writing. Using a comparative judgment model of assessment, four college writing instructors evaluated 85 pairs of essays with one per pair treated by the program and selected the better of each pair. The essays treated by the automated evaluation program significantly underperformed the null hypothesis of 50%. Results suggest the automated evaluation program fails to improve student writing in the eyes of instructors. Theories and implications for why are discussed.

Keywords: formative writing assessment, literacy, composition, writing assessment, pedagogy, automation

Introduction

Teacher feedback on student writing is a cornerstone of writing pedagogy, yet it remains one of the most time-consuming tasks. As a result, both educators and private companies have long sought to leverage technology to make writing feedback more efficient. While the integration of technology into writing assessment is not new, recent developments have shifted the focus to new objectives. Traditionally, Automated Essay Scoring (AES) programs were designed to provide holistic scores for essays. However, these programs have gradually given way to Automated Writing Evaluation (AWE) technologies, which prioritize analytic evaluation over holistic scoring. Whereas AES programs rank and score writing through quantitative and summative assessments, AWE systems aim to edify the writer by providing qualitative and formative instruction. These formative tools go beyond merely scoring; they offer analytical feedback on elements such as rhetoric and style with the goal of improving both the writer and the writing.

The pivot from AES programs, which are intended to score essays, to AWE programs, designed to analyze an essay's component parts—shifting from summative to formative assessment—raises critical questions for writing educators. While AES systems have gained credibility due to numerous studies validating their reliability, particularly in terms of correlating machine-generated scores with those of human raters (Bridgeman et al., 2012; Dikli, 2006), the transition to AWE demands a more nuanced examination of validity. Validation in the context of writing assessment typically involves multiple dimensions, including content, construct, and consequential validity, each of which must be rigorously scrutinized when evaluating AWE programs (Kane, 2013).

Claims about a computer's ability to score an essay along a scale or rank writing according to machine-learned criteria are supported by empirical evidence. For instance, AES programs such as Educational Testing Service's e-rater have demonstrated strong correlations with human scoring in large-scale assessments such as the Graduate Record Examination (GRE), in which machine and human scores reliably correlate at a coefficient between $r = .76$ and $r = .81$ (Monaghan & Bridgeman, 2005). These programs can obtain these correlations because they are trained on vast datasets of human-graded essays, which allows them to approximate human judgment in scoring. However, this form of validation is primarily limited to summative assessment.

In contrast, AWE technology, which seeks to provide not summative but formative feedback—such as comments on structure, coherence, and style—introduces a new set of challenges. Claims about a computer's ability to formatively assess writing and offer qualitative feedback remain largely unvalidated in the literature, reflecting a significant shift in the goals of automated writing assessment. Scholars have questioned whether AWE systems can effectively mirror the depth and complexity of human feedback, particularly when addressing higher-order concerns in writing such as argumentation and voice (Perelman, 2014). This shift from scoring to feedback reveals a need for validation studies that examine not only the accuracy of AWE systems but also their pedagogical effectiveness.

This study contributes to the conversation on validating AWE by focusing on its efficacy as a formative assessment tool rather than on the summative and holistic validation seen in correlational studies of machine and human raters. To begin to validate AWE programs, other metrics beyond score correlations must be considered. For example, construct validity: does the feedback provided by AWE systems align with established writing constructs, or does it focus disproportionately on surface-level issues like grammar and syntax? Additionally, consequential validity must be addressed: what are the broader educational impacts of relying on AWE for formative assessment?

What consequences do educators and students stand to face if AWE is widely adopted in writing pedagogy? Some detractors fear that students might over-rely on automated feedback for revision, potentially bypassing deeper cognitive engagement with their writing, while proponents welcome a method for making education more efficient.

The implications of fully and even semi-automated writing pedagogy are profound, particularly as the automation of other sectors of the economy and society looms large. Given the current advances of generative AI and other technologies used in AWE and text generation, this article attempts to examine the effects of using AWE technology on improving student writing, with a focus on the validity of AWE as a formative rather than summative assessment instrument.

The study described in this article was designed and conducted before the emergence of generative AI technologies like ChatGPT and other large language models. As a result, it does not examine generative AI's potential to offer effective feedback on student writing. However, the study remains valuable for assessing the automation of various aspects of the writing process more generally.

As generative AI continues to take over components of the composition process, scholars and educators will likely focus on identifying which parts of writing remain best served by traditional student-teacher interactions. By exploring the formative feedback capabilities of computers—and noting where automated feedback may falter—this study offers insights into which stages of writing (brainstorming, invention, composing, and revision) might benefit more from human guidance and which could be effectively automated.

Literature Review

Formative Writing Assessment

Non-automated formative writing assessment currently exists in several configurations. Most typically, human teachers provide detailed feedback on student writing, instructing students through a recursive series of drafts, feedback, and revisions. In secondary and higher education, students also commonly receive feedback from peers as well as engage in self-assessment or revision. On many college campuses, writing centers offer one-to-one formative consultations to help students become "better writers" (North, 1984). While AWE is not yet as popular as any of these assessment mechanisms, it is increasingly common as a third-party web-based resource for students to employ at their own discretion.

Scholars debate whether computer-assisted pedagogy can play an effective role in improving writing. Studies analyze its efficacy by comparing it to alternatives (Shermis et al., 2006). In a meta-analysis of the effectiveness of different writing feedback approaches, Graham et al. (2015) compared four feedback mechanisms for student writers in Grades 1–8: adults, peers, self, and computers. They found that all feedback sources resulted in improved writing, but that computers resulted in the smallest positive effect of the group (average weighted effect sizes: adults = .87; peers = .58; self = .62; computers = .38). In a case study examining the use of the Writing Roadmap 2.0 AWE technology for formative feedback, Rich and Wang (2010) also found some positive effects of AWE use on middle and high school writers in the US, with low-performing students in particular seeing the greatest improvement.

Although computer-assisted formative feedback often yields positive effects in experimental research of writing ability, there are some concerns. First, as in the Graham et al. (2015) meta-

analysis, the effect of computers tends to be significantly smaller than human-centered alternatives. Another issue is that most research on this topic involves K-12 education (Landauer et al., 2009; Mao et al., 2018; Myers, 2003); whether a computer program can improve college-level writing remains under-researched. Studies on the effect of general computer-assisted teaching, not just writing instruction, at the college level that do exist have been somewhat inconclusive. In an early meta-analysis, Kulik et al. (1980) showed computer-assisted college instruction made small contributions to course achievement. In an experiment involving students in an online teacher education course, Riedel et al. (2006) found use of AES programs improved the quality of student papers based on final scores by human raters; however, because the study involved only students in an online course, no in-person pedagogy served as a control group, so relative quality remains unknown.

While some K-12 education researchers maintain optimism about the potential applications of AWE technology, composition scholars researching college writing and adult literacy are more critical. Their criticisms center on four primary areas of concern: the theories/constructs of writing undergirding AWE programs; the consequential validity of AWE technology; the repercussions of computerized writing assessment for classrooms, students, and teachers; and the impact of such technologies on underrepresented student populations (Elliot & Klobucar, 2013). To be sure, recent advances in AWE computational models do not mitigate such criticism. As AWE tools have been refined with the integration of Natural Language Processing (NLP), Latent Semantic Analysis (LSA), and generative AI and large language models (LLMs), scholars such as Elliot and Klobucar (2013) remain critical because such advances do not allow the programs to "read" student writing like a human, but rather enable them to produce evaluative comments based on statistical modeling of measurable aspects of the text.

This latter concern—whether machines can really "read" text—represents the primary criticism of AWE programs attempting to evaluate writing beyond assigning numerical scores. Generally, psychometricians and computational linguists view language ability as a cognitive trait and writing as a measurable skill, whereas composition scholars tend to view both language ability and writing as rhetorically complex and socially embedded behaviors resistant to empirical measurement. Strict adherence to the former view, Condon (2013) argues, can result in teachers assigning the kind of writing that can only be evaluated by machines, such as short, timed essays of a specific genre, which proponents of the latter view argue fails to accurately reflect true writing ability. Deane (2013) similarly argues that for AWE programs to succeed in the future, they must be programmed to account for writing's social and cultural dimensions. Otherwise, AWE systems will simply reinforce outdated and narrow notions of formal correctness that fail to account for modern, expanded definitions of the writing construct (Vojak et al., 2011).

Composition scholars also criticize the consequential validity of automated writing assessment—how the data it yields is used and misused. While AES programs are used summatively for scoring essays, critics have long argued that essay scores reflect only a small aspect of the writing construct and are inherently reductive proxy measures for more nuanced writing abilities. In theory, AWE programs are intended to improve on this shortcoming of AES, as they are designed to offer more than a reductive numerical essay score. But the research remains inconclusive, as Balfour (2013) showed that when used as a source of writing feedback for students enrolled in MOOC courses, AWE failed to improve writing compared to having students participate in peer review.

Nevertheless, some believe AWE's pivot to formative assessment shows potential and promise. AWE represents a natural evolution beyond mere summative assessment. Furthermore, AWE technology is believed to offer teachers more time to focus on other aspects of their pedagogy, potentially improving student learning outcomes overall. While there are legitimate concerns that this could lead to the abandonment of traditional writing instruction, proponents argue that AWE technology could simply shift the focus of writing education. If AWE can effectively handle feedback on rough drafts, educators might place greater emphasis on teaching students to become skilled editors rather than solely writers. In this context, students would learn to critically evaluate the suggestions offered by AWE programs. This shift from teaching writing to teaching editing should not be viewed as the abandonment of writing instruction but rather as an adaptation to the inevitable evolution of word processing technology. Just as typewriters and word processors relieved educators from the need to teach penmanship, AWE could streamline the invention and composition processes, allowing the focus of writing and rhetoric to center more on the art of revision, comprehension, and critical thinking.

The primary tension concerning AWE's role in teaching writing therefore involves how much the technology supplements—versus supplants—human teachers. Deane et al. (2013) see a narrow, supplemental role for AWE in large-scale assessment tasks of classroom activities, but they maintain that such tasks must be limited in scope to the measurable and technical aspects of writing. This position—that computers can only ever play a limited role in classroom instruction—appears to be the consensus among rhetoric and composition scholars.

While AWE technology promises efficiency to some in writing assessment, it is crucial to interrogate the broader implications of this efficiency, particularly in the context of academic labor. Efficiency arguments often frame AWE as a tool to reduce the time-intensive aspects of grading and feedback, ostensibly freeing teachers to focus on more meaningful pedagogical work. However, as the recent MLA-CCCC Task Force on Writing and AI (2023) has noted, the widespread adoption of such tools could disproportionately affect precariously employed writing instructors. For contingent faculty, graduate instructors, and other non-tenure-track educators, efficiency imperatives might translate into increased workloads rather than relief. Institutions could justify larger class sizes or additional teaching responsibilities, effectively offsetting any time saved by AWE. This risks perpetuating a cycle where the most vulnerable educators shoulder the brunt of cost-cutting measures, with little tangible benefit to their professional development or pedagogical goals.

Furthermore, the efficiency narrative obscures the complex labor of writing assessment itself. Assessing student writing is not merely a technical task; it is an interpretive and rhetorical act that involves understanding context, intent, and individual learning trajectories; teacher and student are in negotiation, not transaction. By framing AWE as a means to streamline this process, there is a danger of devaluing the nuanced expertise that educators bring to writing assessment. While AWE might offer useful supplemental feedback, it cannot replicate the dialogic and iterative nature of teacher-student interactions that are central to effective writing pedagogy. Thus, the potential for AWE to enhance efficiency must be weighed against the risk of reducing writing instruction to a transactional process, where the labor of assessment becomes increasingly commodified and disconnected from the relational aspects of teaching.

In addition to concerns about the writing construct and the uses (and misuses) of AWE programs, rhetoric and composition scholars have also criticized the administrative and logistical

repercussions of the increased use of AWE technology. Repercussions range in kind, from fundamental changes to classroom dynamics to alterations to curricula to the revamping of course placement and school admissions processes. All manner of writing education stands to be affected by the integration of AWE technology into schools. While some writing program administrators feel pressured to accept the inevitable and strategically adopt the use of AWE programs, others (Cheville, 2004) warn the increasing influence of automated writing educational technology could transform teachers into "data managers" and redefine writing instruction away from a negotiated transaction between writer and reader to a formula to be followed by students.

Finally, there are concerns over questions of identity as they relate to AWE programs. While not yet studied as thoroughly as the above issues, initial studies suggest AWE programs may have "disparate impacts based on gender, ethnicity, nationality, and native language, privileging or penalizing some cultural backgrounds and languages over others" (Elliot & Klobucar, 2013). However, some research regarding AWE use among English language learners (ELL) and students with disabilities, specifically, shows potential benefits for these groups. For example, scholars have theorized that the instantaneous feedback capability of machines is well-suited for L2 and ELL students, pending further validation (Ranalli et al., 2017). In a meta-analysis of 37 studies comparing foreign language teaching supported by AWE technology versus pedagogy not supported by AWE technology, Grgurović et al. (2013) found a small but positive effect of using AWE technology in foreign language teaching. Other studies, however, have proved more inconclusive. Wang (2013, 2015) found that AWE programs improved L2 writing only in certain areas of formative evaluation like error analysis of usage and feedback on organization, and that students preferred a combination of human and machine feedback to automated evaluation alone.

Students with disabilities (SWD) is a group scholars think stands to benefit the most from AWE technology. Wilson (2017) compared growth in writing quality between SWD and typically developing (TD) students when both groups used AWE programs. Results showed a positive association between AWE use and growth in writing quality for SWD, suggesting these technologies may help close achievement gaps in the realm of disability. Overall, rhetoric and composition tends to view AWE programs, along with their uses and consequences, with skepticism, urging educators to approach their adoption cautiously, but they remain interested in specific applications of the technology for certain groups of students.

The Case of Chegg

Most research about AWE tends to involve the proprietary programs of companies like ETS or Pearson. However, as the technology becomes more widespread, we are seeing greater access to AWE programs via online sources. For this reason, I chose to conduct an experiment using an online AWE program, Chegg.com's EasyBib Plus, which is part of one of the most popular online resources for college students. Chegg, known for its online tutoring across all disciplines and textbook rentals, has become a go-to resource for many students, and state-of-the-art writing assistance is one of the company's latest efforts.

Chegg's EasyBib Plus is not a typical freely available online AWE tool. In 2018, Chegg acquired the AI-enhanced WriteLab AWE technology, which had a strong reputation for providing nuanced feedback. WriteLab was developed by a team of computational linguists and English instructors and was respected for its ability to offer detailed, context-aware suggestions (Chegg, 2018; Sternlicht, 2018). After integrating the WriteLab technology, Chegg's EasyBib Plus now flags

a variety of writing issues, including grammar errors, stylistic concerns, punctuation problems, sentence structure, and even higher-order concerns like clarity and argumentation.

The program works as follows. After uploading an essay, the EasyBib Plus then flags various issues across the essay, such as vague wording, passive voice, unclear transitions, sentence fragments and combination, issues of concision, and basic grammatical errors like subject-verb agreement and comma splices. More specifically, the tool underlines sections of text and provides suggestions like, "consider rephrasing for clarity" or "this sentence is passive; revise for active voice" or "consider combining these sentences." Additionally, it identifies areas where the argument could be strengthened, such as weak thesis statements or insufficient evidence to support claims.

The user interacts with Chegg's feedback through an interface that allows for direct revision. When a student hovers over a flagged section, a pop-up appears with a detailed explanation of the issue and suggestions for improvement. Students can then choose to accept the suggestion, revise the text independently, or ignore the feedback. This interactive process is intended to simulate the revision process that would occur with human feedback, encouraging students to engage critically with their writing.

After exploring the EasyBib Plus AWE program, I conducted a mixed-methods quasi-experimental case study to test Chegg's claims and gauge the relative value of using its AWE program to formatively assess student writing. This study aimed to test the capabilities of Chegg's AWE tool, specifically whether it improves persuasive argumentative writing in the eyes of college composition instructors. The study design attempted to isolate the independent variable of Chegg's EasyBib Plus tool and measure its influence on the dependent variable of essay quality, as compared to essay versions unedited by AWE technology. This study serves as an initial exploration of the efficacy of AWE for formative feedback that attempts to imitate what is normally provided by a writing teacher.

Methods

The study that follows is best described as a small *n* quasi-experiment.[1] The experiment tests Chegg's claims that its EasyBib Plus program can improve student writing quality by comparing writing treated by the EasyBib Plus to untreated drafts. Although only the EasyBib Plus program was tested, some findings may apply generally to the automation of writing feedback at the college level independent of specific AWE tools. The goal of this experiment is to analyze an instructional supplement for student writers, one that Chegg, a massively popular college tutoring site, has invested significant money in. Though limited in size, scope, and generalizability, the experiment opens pathways to broader questions about the limitations and applications of AWE and identifies potential pitfalls for automated education more generally.

Participants

Four graduate student English instructors participated as raters. At the time of participation, all instructors had a minimum of three years of experience teaching college English as instructors of record; all had experience teaching their own sections of the first-year composition (FYC) course for which the essays used were written; all were PhD students in the English department at Purdue University; and all had taught, or were currently teaching, some form of argumentative writing in their classes.

1 The research project was IRB approved. The protocol number is 1904021987.

Design

This study design is a within-subjects (or repeated measures) quasi-experiment. While I cannot claim this experiment meets the criteria to be labeled a true experiment, Ary et al. (2014) note that in much educational research, true experiments are impossible due to the ethical limitations of randomization of students: "neither the school system nor the parents would want a researcher to decide to which classrooms students were assigned" (p. 339). Therefore, quasi-experimental designs are often used in educational contexts, which attempt to randomize subjects as much as possible within given curricular constraints. Such is the case in this study; since I had no control over whether the student essays used were a truly random sample, I can only claim that the sample of essays is quasi-random.

The quasi-experimental design involved randomly assigning each of the four raters 25 pairs of student essays from a sample of 85 pairs, 20 of which were unique to their sample and 5 of which were shared and evaluated by all four raters as a control mechanism. Each pair contained two versions of the same essay written by the same student, with one essay per pair treated by the AWE program. The raters were instructed to read each pair of essays sequentially and designate one of the pair "better" to determine the frequency with which the essays evaluated by the AWE program are perceived as better than their untreated counterparts.[2]

The "which is better" method of evaluation draws on L.L. Thurstone's (1994) law of categorical judgment. Thurstone's law is a model used in pairwise comparisons to measure differences in perceptions. It describes a technique commonly found in educational and psychometric research, which attempts to isolate non-physical traits or attitudes of the mind: "the law is applicable . . . to qualitative judgments such as those of excellence of specimens in an educational scale" (Thurstone, 1994, p. 266). A pairwise comparative model is particularly appropriate for this experiment because it does not require the use of an assessment rubric; use of a detailed rubric in this experiment is a threat to its validity, since use of a rubric risks measuring how well the rubric is applied to the evaluation of an essay rather than the determination of which essay is perceived as better.

At nearly every stage of sample preparation, materials were randomized. The 85 essays were randomly selected from a population of 100 students across five different FYC courses. Five essays were randomly selected as a control sample to be evaluated by all four raters; the remaining 80 essays were randomly divided into four groups of 20 and randomly assigned to the four raters. When the raters were given a single stack of printed essays, the order of which essay (treated or untreated) appeared first was also randomized.

The experiment was designed for a point estimate analysis, which is the estimation of an unknown population parameter value. Point estimates are values between 0 and 1 that represent a probability, in this case the probability a Chegg-treated essay is designated "better" than its untreated counterpart. Since this experiment involves only one of many possible samples, a 95% confidence interval was calculated to produce a range of probabilities that the treated essays are perceived better by the raters, a more accurate estimation of the program's true success rate due to inherent sampling error. The range of point estimate values is tested against a null hypothesis value for statistical significance. In this case, the null hypothesis assumed a point estimate value of $\hat{p} = .50$, or 50%; in other words, the null hypothesis assumes that the probability of a treated or

2 The deliberately ambiguous term "better" was used in order to focus the experiment on the program's claims, which are similarly vague. In addition, by using a vague term across four different raters, the experiment better assesses the program's abilities rather than how well the raters would have applied specified criteria.

untreated essay being designated better is equal, like a coin flip. This would mean that the EasyBib Plus program provides no discernible improvement or detriment whatsoever to the writing quality of the essays. If the estimate does differ from $\hat{p} = .50$, we need to know if it was likely due to chance, so the point estimate would then undergo a paired sample t-test to see if it differed significantly from the $\hat{p} = .50$ null value, either positively or negatively.

Materials

The essays used were collected from my own FYC courses, five different sections of FYC (20 students per course, 100 essays total from which 85 were randomly selected) taught between Fall 2015 and Fall 2017. At the time the essays were written, the writers varied by class standing, but most were first- or second-year students majoring in a variety of disciplines. The 85 selected essays thus represent a quasi-random and quasi-representative sample of (early) university students, since FYC is a general education requirement and one of the most widely taken classes on campus.

Furthermore, all essays were taught using the same assignment prompt and rubric and were written at approximately the same point in each semester—as the second assignment in the sequence of four major projects. This consistency provides further control among the samples, since each rater was evaluating papers that were written by students at roughly the same point in the semester, despite them coming from five different semesters. Papers written by students at the end of the semester versus the beginning might vary significantly in quality, introducing unwanted variance into the experimental results. The primary difference between essays from different sections in this sample, then, is only whether they were written in a Fall or Spring semester. All essays were completely de-identified and otherwise unmodified except to match typefaces, font sizes, and spacing. A numerical code in the top left corner was assigned to each essay so only the researcher could identify the treated essays.

Chegg's AWE tool, EasyBib Plus, was used to treat the essays. EasyBib Plus offers instantaneous feedback on issues of style, grammar, writing clarity, and plagiarism (Chegg, 2018). Use of Chegg's EasyBib Plus program required me to manually enter each of the 85 essays into the program and accept grammar and style change suggestions at my own discretion.[3] All essay data was saved in an Excel spreadsheet, and as the essays were treated, certain attributes were logged for later comparison to ensure each of the four samples contained essays of roughly equal quality and length. These attributes included average word counts of both treated and untreated essays, changes in word counts between the treated and untreated essays, number of edits made from the AWE revision suggestions, and ratios of changes in words per edit (see Table 1). These attributes helped determine if re-randomization was needed to ensure each sample was roughly equal to the other and the overall sample.

The final material consideration for the experiment was which kinds of essays to use. Argumentative writing was decided on because it is a genre commonly taught in FYC that often emphasizes issues of style, something many AWE programs claim to address. Editorials were chosen as a specific genre of argumentative writing because they are, again, commonly assigned in writing classes and relatively brief.

3 This retro-treatment of essays is a limitation of the study but was done due to time constraints. A more rigorous study would have students use the program themselves. Even though I can only ever approximate how the program might be used, I attempted to use it consistently and not in a way that would disfavor the program.

Table 1

Essay Attributes for Each Rater Sample and Overall

Rater Sample	Essay Attributes				
	Words, Untreated	Words, Treated	Change in words	Edits	Word change per edit
Sample A	910.05	889.85	-20.20	29.90	.70
Sample B	926.95	908.80	-18.15	29.25	.64
Sample C	919.60	897.70	-21.90	29.25	.77
Sample D	904.50	887.10	-17.40	27.40	.66
Overall	915.28	895.86	-19.42	28.95	.69

Procedures

The experiment was conducted on campus over the course of four weeks in July and August 2019. After raters read and signed a consent form, I provided some basic background information about the assignment for which the essays were written. Careful not to influence how the raters should evaluate the essays, I simply tried to give the raters an idea of what the assignment looked like in class. I explained the assignment was an editorial, and students were encouraged to write about timely topics with no right or wrong answer—to argue a position or offer an opinion on a topic and support it. I explained that students were assessed not only on the coherence of their arguments, but also on the style, clarity, and persuasiveness with which they wrote.

The procedures of the experiment were then explained: In front of each rater was a single stack of 25 pairs of essays (50 total). Their job was to read each pair in the order presented, or to read the pair simultaneously, and designate one of each pair "better." They did not know that one of each pair had been treated by an AWE program, and they did not know that the order of each pair (that is, which essay of the pair appeared first) was randomized between the treated and untreated essays. Each pair was given a letter and numerical code during the de-identification process (for example, A29 and A34, B16 and B12), and after reading each pair, they entered the numerical portion of the code for whichever essay they designated better into a Qualtrics survey. They were allowed to take breaks as needed. Each of the four raters took between 2–3 hours to complete this portion of the experiment.

After each rating, I conducted a brief interview, lasting approximately 5–10 minutes with the following questions.

1. Basic information about instructors:
    a. What year are you?
    b. How many years have you taught college English classes?
    c. Do you teach argumentative writing in your classes?
2. Describe the criteria you used to determine which essay was better.
3. Did your criteria change over time?
4. How confident were you in designating one essay better than the other?

5. How different did you think each pair of essays were from one another?
6. Which features of writing do you think make for a well-written editorial/argumentative essay?
7. What changes do you look for when assessing the quality of revisions from a rough to a final draft?
8. What strategies do you suggest for your students to follow when revising from a rough to final draft?

During these interviews, I tried to get a sense of the criteria each rater used/developed for designating one of the pair better, since I did not enforce any assessment criteria. The purpose of these interviews was to learn how the raters interpreted the differences in the essays, and to determine if they approached the task with significantly different mindsets. These interviews, in addition to the shared sample of 5 five essays that all raters read, served as a mechanism to determine if the results should be analyzed in total ($N = 80$), as if each rater were interchangeable, or independently ($n = 20$), as if each rater represented its own probability. In the interest of full disclosure, I have included the results of both analyses in the section to follow.

Results

The Essay Experiment

The experimental results can be analyzed in two ways. In the first, the analysis considers the total sample overall ($N = 80$) by assuming each of the four raters are interchangeable—that they approached the rating task similarly enough that the majority of potential irrelevant variance is controlled for by experimental design and randomization. The advantage of this mode of analysis is the larger sample size, which provides greater confidence in the results. To gauge the viability of this analysis, all raters read the same five pairs of essays at the beginning of their sorting. This shared sample ($n = 5$), as well as the post-experiment interview during which raters were asked questions to better understand their individual processes and assessment criteria, was analyzed for noticeable differences. In the shared sample, no rater designated Chegg-treated essays better more than 40% of the time, suggesting a similarity in their approaches (see Table 2).

Table 2

Chegg-Treated Essays Designated Better in Shared Sample

Rater	Chegg Essays Rated Better	Percentage of Chegg Essays Rated Better
Rater A	1/5	20%
Rater B	1/5	20%
Rater C	2/5	40%
Rater D	0/5	0%

The second mode of analysis assumes the raters are not interchangeable, and that each rater represents a different probability of the EasyBib Plus program's success at improving writing. Using this method, four separate samples are analyzed (n = 20), each with their own set of results. Although this method greatly lowers the sample size, 20 pairs of papers per rater maintains ecological validity in that the sample size resembles the size of most first-year composition classes, mirroring experimental conditions to those of real-life (Brewer, 2000). The advantage of this method is that it does not assume all raters had the same approach, which the shared sample and the interviews cannot fully confirm. Given the lack of explicit direction—raters were instructed to use the deliberately-ambiguous rubric of "which is better" to assess the pairs of essays—not assuming interchangeability among raters might prove more accurate.[4] The results of both modes of analysis are provided below (see Table 3).

Using the first analytic method, the overall analysis yields a point estimate value of = .30 and a 95% confidence interval (CI) of [.20, .40] for the Chegg-treated essays. In other words, out of 80 pairs of essays, raters on average designated the Chegg-treated essays better 30% of the time (24/80); and if the experiment were repeated 100 times, we could expect 95 of the experiments to yield a point estimate value between .2-.4, or 20–40%. The null hypothesis assumed that the

Table 3

Point Estimate Analysis of Chegg-Treated Essays Designated Better

Rater	Treated	$\hat{p}$	SE	95% CI	t(df)	p
Rater A**	2/20	.10	.07	[-.03[a], .23]	$t(19)$ = -5.96,	< .001
Rater B*	4/20	.20	.09	[.02, .38]	$t(19)$= -3.35	.0016
Rater C	9/20	.45	.11	[.23, .67]	$t(19)$= -.90	p = .19
Rater D	9/20	.45	.11	[.23, .67]	$t(19)$= -.90	p = .19
Overall**	24/80	.30	.05	[.20, .40]	$t(79)$= -3.90	< .001

Note. $\hat{p}$ = p-hat, the probability a treated essay is designated better. SE = standard error. CI = confidence interval.

[a] The lower bound of Rater A's confidence interval is slightly negative; it is acceptable to interpret this value as 0, since the parameter it is estimating is a positive value.

*p < .01. **p < .001.

4 While this method is likely more accurate, all raters share very similar backgrounds and have similar teaching and professional experiences. We can assume their enrollment in the same graduate program and similar training provides comparable assessment approaches.

treated essays would not differ meaningfully from the untreated essays in either improvement or detriment. Therefore, a null hypothesis point estimate value of $\hat{p} = .50$ (50%) was assumed, and results were tested against this value for significance. A paired sample t-test yields a t-statistic of $t(79) = -3.90$, $p < .001$, meaning the $\hat{p} = .30$ value is statistically significant.

Analyzed independently, two raters (C, D) perceived the Chegg-treated essays better at a rate of 45% [.23, .67], slightly worse than a coin flip, which is not a statistically significant difference from the 50% probability of the null hypothesis. Conversely, the other two raters (A, B) perceived the *untreated* essays as significantly better on average, designating the Chegg-treated essays better only 10% [-.03, .23] and 20% [.02, .38] of the time, respectively, which both differ significantly from 50%. Overall, the analysis puts the success of the program at approximately 30%, with the 95% confidence interval suggesting we could be confident that a Chegg-treated essay would on average be perceived as better only between 20–40% of the time compared to an untreated essay. Using either analysis, the AWE program at best achieves a probability of improvement of approximately 45%, roughly equivalent to a coin flip; at worst, the AWE program has significantly worse odds of improving a paper (10%) than leaving the essay unedited.

The Instructor Interviews

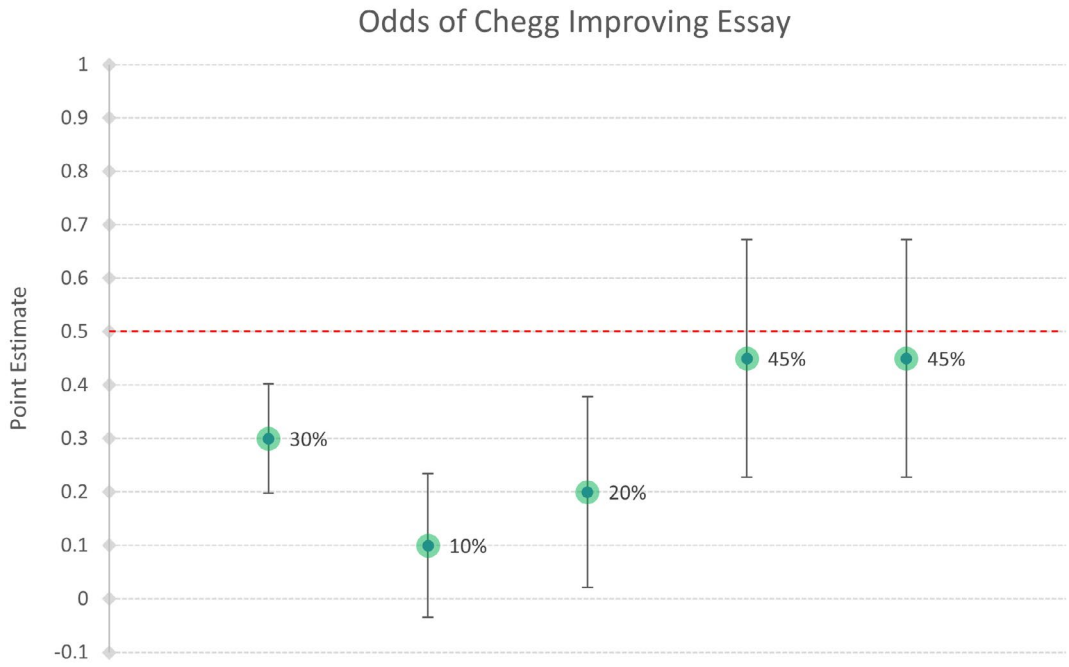
At the conclusion of the experiment, I conducted brief (5–10 minutes) informal interviews with each rater (Blakeslee & Fleischer, 2007). The goal of the interviews was to gain a clearer picture of how the raters approached the task of evaluating the essay pairs, which the quantitative data cannot fully reveal. If the raters varied greatly in their approach, the overall analysis (N = 80) is less valid, since the difference in individual approaches would be the most determinant factor for the experimental results, and an analysis of each sample individually (n = 20) would be required. However, if the raters approached the task similarly enough, they could function as interchangeable in terms of the experiment, validating the overall analysis. In addition, the interviews help to provide more clarity about the effect of the EasyBib Plus AWE program and how the raters perceived the changes it made to the writing under evaluation.

As discussed in the participants section, each of the four rater's backgrounds were extremely similar, and so their response to question one was approximately the same. Questions two and three saw some variation. Since a categorical judgment model of assessment was used, raters were free to draw on whichever criteria they saw fit to determine the better of the two essays per pair. For example, rater A described using criteria based on essay features such as word choice, syntax, and the overall "flow" of the language. Rater B focused their criteria on which essay communicated its meaning—as determined by the rater—more successfully. All raters stuck to their criteria throughout the experiment, with little change throughout.

Questions four and five prompted similar responses from all raters. Raters thought that, overall, the differences between essays in each pair were not great, which somewhat lowered their confidence in designating one of each pair better. Each of the raters described the differences as something like "surface-level" or at the level of "lower-order" concerns. Despite the agreement that the differences were not great, each of the four raters, as well as each of their practice samples, yielded a Chegg success rate below 50%. Future research might be interested in the disconnect between these interview responses and the quantitative data yielded by the experiment—as the interview data would seem to predict each of the four raters would yield point estimates closer to the null value like those of raters C and D (see Figure 1).

Figure 1

Odds of Chegg Improving Essay



Note. *p < .01. **p < .001

Questions six, seven, and eight prompted mixed responses, but each were relatively similar. These questions were designed to spur a discussion of genre and revision pedagogy. Rater D commented on the more informal tone of editorial writing, and others echoed that sentiment. All raters discussed that the argument should be clear and persuasive in an editorial, and that effective editorials are ones that know their target audience. For revision pedagogy, rater C made a distinction between sentence-level revision and "big picture" revision—such as that of thesis statements and entire paragraphs—and seemed to suggest that the differences in essays were primarily noticeable at the sentence level.

In sum, the interviews helped to contextualize the quantitative data yielded by the experiment. While the responses to questions about how different the essays were and the raters' confidence in designating one better than the other somewhat contradict the quantitative data, the interviews also give greater insight into the teaching of editorial writing at the college level and potentially why the Chegg EasyBib Plus tool failed to improve student writing. There appears to be broad consensus on the characteristics of the genre, as well as an understanding of the difference between higher- and lower-order concerns. Nonetheless, the greater volatility in the analysis of the raters individually could reflect the variation inherent to different classes, semester by semester, since each of the raters' samples (n = 20) is approximately the same size as a typical FYC course.

Naturally, when several classes are combined, the variation is minimized, as the overall sample (N = 80) reflects.

Discussion

The experiment found that Chegg's EasyBib Plus AWE tool was unsuccessful at improving editorial essays as evaluated by four experienced college English instructors. The experiment's null hypothesis assumed the probability of designating a Chegg-treated essay "better" to be approximately 50%. In reality, on average, the 80 Chegg-treated essays had only between a 20–40% probability of being designated better than their 80 untreated counterparts, a statistically significant difference in a negative direction from the null hypothesis value. This suggests features of the program itself, and not random chance, were responsible for rendering the treated essays worse than their unedited versions. The results should be interpreted cautiously with regards to generalizability. Nonetheless, the results open an interesting discussion about the inherent limitations of AWE technology to formatively evaluate writing, specifically AWE's inability to parse genre and meaningfully address higher order writing concerns at the college level. In what follows, I will discuss some initial theories regarding the results.

AWE and Genre

A major obstacle for AWE programs is genre. AWE programs have traditionally been used in summative evaluation efforts such as scoring written placement exams or standardized test essays, where test takers submit an essay written in a very narrow genre whose criteria the AWE programs have been trained on. Summative *scoring* of essays by machines correlates very highly with those of human raters (Dikli, 2006). But scoring and improving essays are very different tasks. Moreover, without the narrow genre constraints provided by a standardized essay exam, AWE programs may struggle to comprehend the larger context within which the writing has occurred, which in the case of this experiment is an argumentative editorial. The challenge of parsing genre, as well as a closer analysis of what the EasyBib Plus tool actually does, can help explain the results.

When a user logs onto Chegg to access the EasyBib Plus tool, they are prompted with a submission portal that asks for no genre information. The user uploads or copy-pastes a paper into the portal, and then the program provides suggestions for revision. There are no questions about genre or assignment guidelines; the EasyBib Plus seems to view writing as simply writing, independent of genre constraints or context. This acontextuality is likely a contributor to the tool's inferior performance in the experiment. Without knowledge of the assignment's genre or context, the tool is limited to providing mostly general writing advice, which may or may not be applicable to a given genre.

For example, a common suggestion the tool made was to omit intensifier words like "very" or "just." Elimination of such words may help with an essay's formality or tone, but in the context of an editorial could temper the forcefulness or obscure the clarity of the argument or writer's position, thus weakening the quality of editorial writing specifically. The program offered many other surface-level suggestions similarly aimed at tone or general writing "improvement" that could have conflicted with expected genre characteristics. Because I was careful to let the raters use their own criteria to define "better" however they saw fit, and since the findings were fairly consistent across four different raters, it seems reasonable to assume the program was unable to navigate the editorial genre appropriately.

"Teaching to the Test" and "Writing to the Program"

Another issue to consider is the flattening of 80 individual writing voices that results from feeding each essay through the same AWE program. Because the EasyBib Plus is insensitive to genre and therefore offers only general writing advice, many of the suggestions it made were repeated across essays. Although many of these suggestions were simple word or phrase adjustments, the effect produced a kind of uniform writing voice, which is likely amplified to raters reading 25 pairs of essays back-to-back.

The flattening of writing voices by AWE algorithms recalls an insight from social scientist Donald Campbell (1979). Writing about research methodology in the 1970s, Campbell formulated what would later become known as Campbell's law, which states, "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (p.85). In other words, Campbell is talking about the conflation of measurement and achievement, when a measure becomes a target, which, for our purposes, is similar to the conflation of summative and formative evaluation mechanisms.

Campbell's law is popularly used to explain the phenomenon of "teaching to the test," which occurs when summative measures like standardized test scores are inappropriately used for accountability, through which they become corrupted and used erroneously as formative educational tools. As Campbell (1979) explains,

> achievement tests may well be valuable indicators of general school achievement *under conditions of normal teaching aimed at general competence*. But when test scores become the goal of the teaching process, they both lose their value as indicators of educational status and distort the educational process in undesirable ways. (p. 85)

Just as the overvaluing of standardized test scores as evidence of academic achievement results in teachers "teaching to the test," an overreliance on AWE programs for formative writing pedagogy could result in students "writing to the program." Just as students can score highly on a test without learning anything of substance beyond test-taking techniques, so too could students learn to satisfy the dictates of lower order writing conventions valued by AWE algorithms without learning anything of substance about writing itself.

Proponents of AWE technology often argue that the time freed by using such technology in place of teacher feedback could result in improvements to courses overall, as teachers would have more time to focus on other class components. The results of this experiment suggest that not only would automated writing feedback not improve student writing, but it might actually worsen it, or at the very least result in more uniform, but not necessarily better, writing voices.

Limitations

This experiment is limited in many ways. Although a quasi-representative and random sample of college students was obtained, 85 is a small sample size. A larger sample of student essays, as well as a larger number of raters, would provide greater confidence in the results. In addition, only one AWE program—Chegg's EasyBib Plus—is tested. Chegg is an extremely popular service, especially among college students, and while its AWE tool is comparable to similarly accessible online resources, other AWE programs might perform differently given the same experimental conditions. More research is needed about other programs, especially regarding AWE tools like

the Linguistic Inquiry and Word Count (LIWC) that claim to feature genre-sensitive evaluation of writing.

Another significant limitation in the study's design is the retro-treatment of student essays. Because I fed each essay into the EasyBib Plus tool myself and accepted revision suggestions at my own discretion, rather than student writers performing this task on their own, the actual changes made to the essays represent an approximation of how the program might be used. I tried to be consistent in the changes I accepted, and carefully logged as many essay attributes as I could to monitor my consistency, and I tried not to deliberately accept changes that would make the program look worse. During this process, I accepted the program's suggestions as often, but also as realistically, as possible. For instance, I tended not to accept any change that would too greatly alter or obscure the original meaning of the text. I tended not to accept changes that appeared to be blatant errors, like changing verb tenses from what naturally sounds correct. For the suggestions that required judgment calls rather than accepting simple yes/no revisions, I attempted to insert myself into the essay as little as possible to resolve the identified error; for example, the program frequently suggested resolving sentence fragments by "adding in missing information" or "combining one sentence with a nearby sentence," and I always tried initially for sentence combination.

This element of the experiment is imperfect, as it is impossible to replicate how individual students might use such a program in their own way. However, I tried to engage with the program as honestly and consistently as possible. But I also accepted as many revision suggestions as I could, hoping to test the program's true abilities, which may reflect how a great number of students use such technology. Additionally, due to this quasi-experiment's particular methodology, the results of this experiment only allow us to speculate on whether the writing, and not the writer, was improved by the AWE program. In the future, research like this would be stronger if students used the AWE program themselves to capture the variation in approaches to using the program more accurately. Additionally, a follow up study involving a separate measure of writing ability could be employed to triangulate and compare data on the improvement of student writing ability more generally.

Conclusion

This experiment suggests that AWE programs have inherent limitations in significantly improving writing quality at the college level. The program's poor performance may come as a surprise—or a letdown—to educators interested in streamlining feedback for student writing. Although this experiment has its limitations and cannot definitively confirm theories, the data supports informed speculation and skepticism regarding the ability of AWE programs to provide formative evaluation and qualitative feedback on college student writing. The EasyBib Plus program claims to enhance grammar, clarity, and writing style, enabling students to submit their "best paper." However, this study suggests that using Chegg's program results in a "better" paper than an unedited first draft only 20–40% of the time—a significant negative deviation from the null hypothesis. While further research is necessary, current claims about the formative educational value of AWE should be approached with caution.

References

Ary, D., Jacobs, L. C., Sorensen, C., & Walker, D. A. (2014). *Introduction to research in education*. Cengage.

Balfour, S. P. (2013). Assessing writing in MOOCs: Automated essay scoring and calibrated peer review. *Research and Practice in Assessment, 8*, 40–48.

Blakeslee, A., & Fleischer, C. (2007). *Becoming a writing researcher*. Lawrence Earlbaum Associates.

Brewer, M. (2000). Research design and issues of validity. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 11–26). Cambridge University Press.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*(1), 27–40. https://doi.org/10.1080/08957347.2012.635502

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*(1), 67–90. doi:10.1016/0149-7189(79)90048-X

Chegg. (2018). *Chegg deepens investment in writing and ai with acquisition of WriteLab*. Retrieved August 28, 2019, from https://investor.chegg.com/Press-Releases/press-release-details/2018/Chegg-Deepens-Investment-In-Writing-And-AI-With-Acquisition-Of-WriteLab/

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *The English Journal, 93*(4), 47–52.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays. *Assessing Writing, 18*(1), 100–108.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7–24.

Deane, P., Williams, F., Weng, V., & Trapani, C. S. (2013). Automated essay scoring in innovative assessments of writing from sources. *Journal of Writing Assessment, 6*(1). https://escholarship.org/uc/item/3nf6r4kv

Dikli, S. (2006). An Overview of Automated Scoring of Essays. The Journal of Technology, *Learning and Assessment, 5*(1). https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1640

Elliot, N., & Klobucar, A. (2013). Automated essay evaluation and the teaching of writing. In M. D. Shermis & J. Burnstein (Eds.), *Handbook of automated essay evaluation* (pp. 16–35). Routledge.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: a meta-analysis. *The Elementary School Journal, 115*(4), 523–547. https://doi.org/10.1086/681947

Grgurović, M., Chapelle, C. A., & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL, 25*, 165–198. https://doi.org/10.1017/S0958344013000013

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review, 42*(4), 448–457.

Kulik, J. A., Kulik, C.-L. C., & Cohen, P. A. (1980). Effectiveness of computer-based college teaching: a meta-analysis of findings. *Review of Educational Research, 50*(4), 525–544. https://doi.org/10.2307/1170294

Landauer, T. K., Lochbaum, K. E., & Dooley, S. (2009). A new formative assessment technology for reading and writing. *Theory Into Practice, 48*(1), 44–52. https://doi.org/10.1080/00405840802577593

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H.-S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment, 23*(2), 121–138. https://doi.org/10.1080/10627197.2018.1427570

MLA-CCCC Task Force on Writing and AI. (2023, July). *MLA-CCCC Task Force on Writing and AI working paper: Overview of the issues, statement of principles, and recommendations*. Modern Language Association of America; Conference on College Composition and Communication.https://hcommons.org/app/uploads/sites/1003160/2023/07/MLA-CCCC-Joint-Task-Force-on-Writing-and-AI-Working-Paper-1.pdf

Monaghan, W., & Bridgeman, B. (2005). E-rater as a quality control on human scores. *Educational Testing Service*, 1–4.

Myers, M. (2003). What can computers and AES contribute to a K-12 writing program? In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross- disciplinary perspective* (pp. 3–20). Lawrence Erlbaum Associates.

North, S. M. (1984). The idea of a writing center. *College English, 46*(5), 433–446. https://doi.org/10.2307/377047

Perelman, L. (2014). When 'the state of the art' is counting words. *Assessing Writing*, *21*, 104–111.

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, 37(1), 8–25. https://doi.org/10.1080/01443410.2015.1136407

Rich, C. S., & Wang, Y. (2010). Online formative assessment using automated essay scoring technology in China and U.S.—Two case studies. *2010 2nd International Conference on Education Technology and Computer, 3*, V3-524-V3-528. https://doi.org/10.1109/ICETC.2010.5529485

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. https://doi.org/10.2190/U552-M54Q-5771-M677

Shermis, M. D., Burstein, J., & Leacock, C. (2006). Applications of computers in assessment and analysis of writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 403–416). Guilford Publications.

Sternlicht, A. (2018, May 25). His company WriteLab was acquired by Chegg before he turned 30. *Forbes*. Retrieved September 14, 2019, from https://www.forbes.com/sites/alexandrasternlicht/2018/05/25/his-company-writelab-was-acquired-by-chegg-before-he-turned-30/

Thurstone, L. L. (1994). A law of comparative judgment. *Psychological Review, 101*(2), 266–270.

Vojak, C., Kline, S., Cope, B., McCarthey, S., & Kalantzis, M. (2011). New spaces and old places: An analysis of writing assessment software. *Computers and Composition, 28*(2), 97–111.

Wang, P. (2013). Can automated writing evaluation programs help students improve their English writing? International *Journal of Applied Linguistics & English Literature, 2*(1), 6–12. https://doi.org/10.7575/ijalel.v.2n.1p.6

Wang, P. (2015). Effects of an automated writing evaluation program: Student experiences and perceptions. *Electronic Journal of Foreign Language Teaching, 12*(1).

Wilson, J. (2017). 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. https://doi.org/10.1007/s11145-016-9695-z