Comparing the Quality of Human and ChatGPT Feedback on Students' Writing

Structured Abstract:

Background. Offering students formative feedback on drafts of their writing is an effective way to facilitate writing development. Recent advances in AI (i.e., ChatGPT) may function as an automated writing evaluation tool, increasing the amount of feedback students receive and diminishing the burden on teachers to provide frequent feedback to large classes.

Aims. This study examined the ability of generative AI (i.e., ChatGPT) to provide formative feedback on students' compositions. We compared the quality of human and AI feedback by scoring the feedback each provided on secondary student essays on five measures of feedback quality: the degree to which feedback (a) was criteria-based, (b) provided clear directions for improvement, (c) was accurate, (d) prioritized essential features, and (e) used a supportive tone.

Sample. 200 pieces of human-generated formative feedback and 200 pieces of AI-generated formative feedback for the same essays.

Methods. We examined whether ChatGPT and human feedback differed in quality for the whole sample, for compositions that differed in overall quality, and for native English speakers and English learners by comparing descriptive statistics and effect sizes.

Results. Human raters were slightly better at providing high-quality feedback to students in all categories other than criteria-based. AI and humans showed differences in feedback based on essay quality. Feedback did not vary by language status for humans or AI.

Conclusion. Well-trained educators provide higher quality feedback than the current free version of ChatGPT. Considering the ease of generating feedback through ChatGPT and its overall quality, generative AI may be useful in some writing instruction contexts, particularly in formative early drafts or situations when a well-trained educator is unavailable.

Keywords: Automated writing evaluation; formative feedback; writing instruction, generative AI

Introduction

Providing students with formative feedback during the writing process is a key instructional practice that helps students improve as writers (Author, 2011; MacArthur, 2016). By clearly communicating to students what quality performance looks like and how to achieve such performance, formative feedback directs a student toward productive action or improvement in specific writing skills (Author, 2012, 2016; Panadero & Jonsson, 2013; Parr &

Timperley, 2010). However, the considerable time and effort it takes to provide students with feedback, especially multiple students across multiple classes, is daunting for many educators and even deters some teachers from providing needed writing instruction (Applebee & Langer, 2011; Author, 2009; Author, 2019). Therefore, diminishing the burden on teachers to be the sole providers of feedback may create more opportunities for writing and writing instruction.

Automated writing evaluation (AWE) has been studied for years as a way to provide timely evaluation of student writing and diminish the burden on educators to evaluate writing (Wilson & Roscoe, 2020; Wilson et al., 2022). These systems typically use natural language processing and artificial intelligence to evaluate writing. Some studies have shown such systems can produce positive effects on student engagement, efficacy, or writing length and quality (Author, 2010; Author, 2015; Roscoe et al., 2017; Stevenson & Phakiti, 2014; Wilson & Czik, 2016; Wilson & MacArthur, in press; Zhai & Ma, 2022).

While some AWE systems are as efficient and reliable as human raters when assigning scores to writing, they are typically less accurate, more generic and verbose, and sometimes confusing to feedback recipients (Author, 2010; Shermis & Wilson, in press; Wilson & MacArthur, in press). Additionally, preparing such tools for educational settings takes substantial time (Author, 2006; Chen et al., 2022; Moore & MacArthur, 2016; Shermis, 2014; Wang et al., 2020). Historically, AWE systems have required training on hundreds of essays written to the same prompt and iterative calibration with human-provided feedback. These requirements increase their cost and reduce teachers' flexibility in using these systems because evaluation is limited to the types of writing prompts used for training.

However, new forms of generative AI, such as ChatGPT, function differently than previous iterations of AWE software and older AI systems. ChatGPT does not require training on human corpora for a specific task or genre and, for now, is low-cost and accessible. It is possible that new, generative AI like ChatGPT can provide feedback that is timely, targeted, adaptive, and useful—all qualities that help students improve writing (Biber et al., 2011; Author, 2015).

Given the potential promise of generative AI as a producer of formative feedback, we designed this study to examine whether ChatGPT could provide feedback that was similar in quality to human raters, and therefore, potentially useful for process-based writing instruction. We examined the feedback provided by ChatGPT and by human evaluators to the same essay corpus written by middle and high school students in history classrooms in the western United States. We evaluated the feedback based on five aspects of quality emphasized in the extant literature: to what extent the feedback (a) was criteria-based, (b) provided clear directions for improvement, (c) was accurate, (d) prioritized essential features of writing, and (e) was delivered in a supportive tone.

Theoretical Framework

The theoretical framework that served as the foundation for this investigation was the Writer(s)-Within-Community Model (WWC; Author, 2018b, 2023). According to this model, writing and writing instruction are shaped and bound by the communities or contexts in which they take place as well as the cognitive capabilities and resources writers and teachers bring to the task of writing [instruction] in these communities. Writers accomplish the task of writing using five production processes: *conceptualization* (creating a mental representation of the task), *ideation* (generating content from memory or external sources), *translation* (transforming content into sentences that convey intended meanings), *transcription* (transcribing printed or digital text), and *reconceptualization* (engaging in revision).

The current study focused on *reconceptualization* as a student's reconceptualization of the text they produce can be improved by providing quality formative feedback. Such feedback helps a student improve their writing by communicating how a writer's performance compares to ideal writing in a genre and by clearly identifying specific steps a writer can take to improve it (Author, 2018a; Hattie & Timperley, 2007; Hillocks, 1986; Kluger & DeNisi, 1996). Though a writer draws upon multiple inputs when revising, including peer feedback or a writer's own evaluation of their text (Author, 2015), in this study, we compared feedback from ChatGPT and trained human evaluators. Comparing ChatGPT feedback to feedback from trained adult evaluators provided a stronger test of ChatGPT feedback than would be obtained by comparing it to peer- or self-feedback because secondary students are still developing competence as writers and can produce unreliable feedback (Van Steendam et al., 2010).

Because the contexts within which students learn to write differ in multiple ways (e.g., purposes for writing, values placed on writing, writing norms, social practices, and typical practices for writing [instruction]; Author, 2018a), the experiences of fluent English speakers and English learners in various writing communities influence the quality of these students' writing as well as the quality of what weaker and stronger writers produce (Camping et al., 2020; MacArthur et al., 2004). As a result, it is important to examine whether formative feedback differs from the feedback of trained evaluators for writing produced by these groups of writers.

Finally, while effective writing instruction ideally involves teachers frequently providing individual feedback on multiple drafts for each student, such an undertaking is extraordinarily time-consuming (Author, 2006). Despite the empirical grounding of this instructional recommendation (Author, 2018a), process writing coupled with individualized feedback is an infrequent experience for secondary school students in the United States (Applebee & Langer, 2011; Author, 2009; Lawrence et al., 2013). This limits learning, as students learn best when receiving timely feedback (e.g., Black & William, 1998, 2009). Thus, it is essential to examine if alternative approaches (i.e., ChatGPT) for providing secondary students with feedback on their writing are comparable or even superior to more traditional forms of feedback.

Characteristics of Good Feedback

To compare the quality of feedback from ChatGPT and trained human scorers, we assessed the feedback provided by both feedback sources using five specific criteria: the degree to which the feedback (a) was criteria-based, (b) provided clear directions for improvement, (c) was accurate, (d) prioritized essential features of writing, and (e) used a supportive tone. Though providing feedback in a *timely* manner is considered a characteristic of good feedback (Hattie & Timperley, 2007), we did not include this as one of the evaluation criteria in this study. While teachers may or may not provide timely feedback based on numerous contextual constraints, ChatGPT can give feedback immediately and iteratively. Therefore, ChatGPT's feedback is presumably always timelier.

Criteria-Based

Explicitly referencing the criteria in rubrics or standards is a key characteristic of effective feedback (Black & William, 2009; Author, 2018a; Hillocks, 1986; MacArthur, 2016; Roscoe et al., 2013). Connecting feedback to specific criteria makes visible the standards to which students can compare their writing; therefore, students can better understand their progress toward successful writing in a genre (Author, 2018a; Author, 2015; Hattie & Timperley, 2007; Parr & Timperley, 2010). The importance of explicitly referencing criteria to improve student writing is emphasized by studies that find writing rubrics are most effective when they are used to communicate clear expectations and *criteria-based* definitions for good writing (Andrade &

Du, 2005, Panadero, & Jonsson, 2013). However, providing criteria-based feedback can be challenging if teachers or AI evaluators lack the pedagogical content knowledge and understanding of genre conventions as these are needed to articulate what quality performance looks like and how the current text can be improved (Parr & Timperley, 2010).

Clarity of Directions for Improvement

Research indicates that effective feedback is clear and uses precise language to provoke actionable writing strategies (Beach & Friedrich, 2006; Roscoe et al., 2013; Wilson & Czik, 2016). Conversely, vague feedback is less likely to be understood and taken up by writers, especially if a student already has low self-efficacy for writing (American Psychological Association, 2015; Author, 2018a; Hattie & Timperley, 2007; Ranalli, 2018; Roscoe et al., 2013; Wilson & Czik, 2016; Zhu et al., 2020). Ideally, specific steps for improvement are written in precise terms, are easy to follow, and provide scaffolding throughout the writing process.

Accurate

Studies examining the quality of formative feedback emphasize the importance of accuracy as inaccurate feedback can lead to confusion or disengagement during the revision process (Author, 2010; Bai & Hu, 2017; Kluger & DeNisi, 1998; Moore & MacArthur, 2016). For example, in a study examining student perceptions of AWE in the classroom, Authors (2010) argued that inaccurate feedback from AWE resulted in human experts needing to redirect students to specific feedback to focus on.

Prioritization of Essential Features

Because too much feedback can be overwhelming, effective feedback prioritizes essential features of writing that are attainable and reasonable for a student to focus on next (Author, 2010; Black & William, 2009; Underwood & Tregido, 2006). Feedback that addresses writing strategies that a student does not know about or that are inessential will deter them from pragmatic improvement (American Psychological Association, 2015; Author, 2010; Hattie & Timperley, 2007; Kluger & DeNisi, 1998). In contrast, feedback that prioritizes essential features and focuses on higher-order concerns is more likely to support learning, especially if it is manageable, both in terms of quantity and the student's writing ability (Clare et al., 2000; Author, 2018a; Hattie & Timperley, 2007; Underwood & Tregido, 2006; Van Steendam et al., 2010).

Supportive Tone

Finally, effective feedback is affirming, uses a supportive tone, and is non-directive (American Psychological Association, 2015; Author, 2015; Cho & MacArthur, 2010; Hattie & Timperley, 2007; Kluger & DeNisi, 1998; Underwood & Tregido, 2006; Wilson & Czik, 2016). A recent study by Motz and colleagues (2021) found that occasional simple messages praising students for turning in an assignment improved future submission rates and course performance, suggesting the importance of positive feedback for changing behavior. Students' self-efficacy may also affect how they react to feedback written in an [un]supportive tone, with some students less able to productively use feedback perceived as negative (American Psychological Association, 2015; Hattie & Timperley, 2007).

Present Study

This study addressed the following three research questions:

Research Question 1 (RQ1): How does the quality of formative feedback provided by ChatGPT compare to the quality of formative feedback provided by human evaluators?

Research Question 2 (RQ2): Does the quality of formative feedback provided by ChatGPT or human evaluators differ for essays judged as low, medium, and high quality?

Research Question 3 (RQ3): Does the quality of formative feedback provided by ChatGPT or human evaluators differ for English-speaking and EL students?

We did not make specific predictions about the comparisons made between ChatGPT and human scorers in this study because of the newness of ChatGPT and the lack of existing data on its effectiveness for providing formative feedback on writing.

Method

Study Context

Essays from a larger writing intervention were used as the basis for all formative feedback in this study. Students in 26 different classrooms (Grades 6-12) from two school districts in Southern California wrote source-based argument essays in history. Human raters provided written formative feedback on the writing of all students in the study as part of the intervention. For this study, we randomly sampled 200 students from the larger study: 50 students who were designated by their districts as English Learners (ELs); 50 students who were designated as Reclassified Fluent English Proficient (RFEP); and 100 students who were classified as fluent English speakers (IFEP or EO). We used this blocking to provide sufficient power to answer the second research question about the heterogeneity of the quality of feedback provided to EL students. Students were randomly sampled across all grades (6-12) in order to determine the quality of feedback for a broad range of writing in secondary grades. Neither humans nor ChatGPT knew the grade of the student writing and instead focused on the quality of the writing relative to the criteria outlined in the directions provided to them.

Writing Corpus

Across two 50-minute class periods, students wrote to one of two prompts: *How did the Delano Grape Strike and Boycott succeed* or *How did the Montgomery Bus Boycott succeed?*

The prompts required students to read four primary and secondary sources and write an argument of causal analysis. Both prompts emphasized constructing interpretations of the past using evidence and reasoning and required writing skills emphasized in district and Common Core State Standards (Breakstone et al., 2013; CCSI, 2012; Goldman et al., 2016). The prompt, *How did the Montgomery Bus Boycott succeed?*, was adapted from a lesson created by the Stanford History Education Group (Stanford History Education Group, n.d.). Full prompts are included in Appendix A. Student writing was not corrected for grammar or spelling prior to evaluation by humans or ChatGPT so the study would most closely emulate authentic feedback situations in process-based writing.

Feedback Generation

Human Evaluators

The original intervention team recruited sixteen experienced secondary educators teaching social studies or English Language Arts, writing researchers, and graduate students majoring in literacy education to attend a 3-hour training on providing formative feedback. All evaluators had experience teaching and providing feedback to student writing, with 12 evaluators having over 15 years of experience teaching writing. Half of the evaluators were also trained in writing instruction beyond their undergraduate or graduate preparation (e.g., participated in training at their local National Writing Project sites).

The training session described the prompts students wrote to, facilitated a discussion of the evaluation criteria to reference for feedback, instructed scorers on how to locate actionable areas for improvement, and provided guidelines and practice on how to write effective feedback (e.g., *glow and grow strategy*: affirm a specific component of student writing before identifying a particular area for improvement). Human raters had access to rubrics and tips for providing feedback on specific criteria, including (a) content and ideas, (b) evidence use, (c) structure, (d) language use and conventions, and (e) historical thinking.

Researchers asked evaluators to focus on giving feedback on the writing components that would help students make the most growth as writers. Specific sentence stems for feedback were provided. For example, they offered the following response for an essay lacking evidence: “Next time, you might add evidence from the sources to support your claim. Here’s a sentence starter I find helpful: According to the [Source], ‘________.’ This will help you support your claim with clear evidence directly from the source.”

Directions for human raters can be found in Appendix B. On average, evaluators provided feedback for writing in two classrooms (approximately 40-50 pieces of writing). Evaluators had two weeks to provide feedback and received compensation for their work. Providing feedback took between 6 and 8 hours, plus an additional 3 hours of training time. Thus, feedback required approximately 20-25 min of total rater time per essay.

ChatGPT

For the present study, we used ChatGPT (v.3.5) to generate formative feedback for the student writing for which we already had human feedback. To determine the best prompt to elicit feedback from ChatGPT, we used multiple cycles of prompting and feedback analysis by researchers specializing in writing research and technology. Ultimately, we found it helpful to (1) have ChatGPT roleplay as a secondary school teacher to set an appropriate language level and use “a friendly and encouraging tone,” (2) ask ChatGPT to provide 2-3 pieces of specific and actionable feedback to match the instructions given to the human scorers, (3) provide the prompt and sources used by students to respond to the essential question (e.g., "Write an argument that responds to the following question: Why did the Montgomery Bus Boycott succeed?"). The exact prompts used for ChatGPT are available in Appendix C.

The instructions provided to ChatGPT were as follows:

> “Pretend you are a secondary school teacher. Provide 2-3 pieces of specific, actionable feedback on each of the following essays…that highlight what the student has done well and what they could improve on. Use a friendly and encouraging tone. If needed, provide examples of how the student could improve the essay.”

ChatGPT allows users to select a “temperature” setting to adjust the underlying algorithm to be more or less creative. We used a low temperature setting of .1 to reduce the randomness and creativity of the feedback. We dropped two essays from the sample because ChatGPT could not provide any feedback due to the brevity of the essay (less than a sentence). Although human raters received specific suggestions for providing feedback for these extremely brief essays, we did not give ChatGPT similar suggestions because we did not anticipate its failure to respond.

Deductive Coding of Feedback

To compare the quality of human and ChatGPT feedback, we trained research personnel to code the feedback from humans and ChatGPT. For each student (n=200), raters received a single document containing the student essay and feedback from two blinded reviewers—ChatGPT and a human evaluator. Although we randomly ordered whether human or ChatGPT feedback was first or second on the page, the language and tone of ChatGPT feedback were often notably distinct. Also, some human raters provided textual, in-line comments and these were retained. Though it was commonly apparent which evaluator was ChatGPT, placing feedback directly below the essay and having feedback directed to the student made the *experience* of scoring both pieces of feedback similar.

The research team developed and applied a deductive framework to score human- and ChatGPT-generated feedback across the five components of quality formative feedback. These components were determined by a literature review and consultation with three subject-matter experts in the field of secondary writing instruction. Raters scored each feedback component on a scale of 1 to 5, with 1 representing no evidence of quality feedback and 5 representing high-quality feedback for that component. The first component, *criteria-based,* measured how well feedback explicitly referenced criteria for source-based argumentative writing in history. Raters scored this component on a range from (1) all feedback is generic (e.g., good job!) to (5) the feedback consistently and explicitly references criteria for the genre.

The second component, *clear directions for improvement,* measured how clear and actionable feedback was. Raters scored this component on a range from (1) vague (e.g., "where is the evidence?"*)* to (5) consistently provided clear directions for improvement (e.g., "I marked a place where you can add more evidence to support your claim. For example, you might mention the description of __ in Source 2."). Next, *accurate* measured the correctness of each piece of feedback, ranging from (1) completely inaccurate to (5) completely accurate.

Prioritization of essential features measured whether the feedback was essential and attainable based on the student's writing ability as manifested in their essay. Feedback ranged in quality from (1) nonessential to (5) essential and attainable. Nonessential feedback included comments like: "Consider addressing a counterargument to improve the essay" when the essay lacked a clear main claim.

Finally, *supportive tone* was scored on a range of (1) unsupportive to (5) affirming and supportive. Lower-scoring feedback used commanding language and lacked a balance of critiques and affirmations (e.g., "You need to add evidence here" and "This was confusing;

Rewrite”). When providing a critique, the feedback received a higher score if it used suggestive or facilitative language, given the positive impact of non-directive feedback (e.g., “One way you could make this paragraph even stronger is by including evidence from a specific source”). The final codebook is presented in Table 1.

Table 1

Quality of Formative Feedback

Code and description	Score category
Criteria-based Feedback should explicitly reference criteria for quality source-based argument writing in history Quality of feedback ranges from **does not explicitly reference criteria (generic) → explicitly references criteria** **Criteria:** claim, evidence-based, argument, reasoning to support argument/claims **Generic:** summarize, details, information	**5:** *Consistently* explicitly references criteria of source-based argumentative writing (SBAW) in history; minimal to no feedback is generic
	4: *Most* feedback explicitly references criteria for SBAW in history, but *some* feedback is generic/does not explicitly reference criteria in this genre; references to criteria are less explicit than a 5
	3: *Half of the* feedback explicitly references criteria and *half* is generic
	2: Most feedback is *generic*; one piece of feedback somewhat explicitly references criteria for source-based argument writing
	1: *Does* not explicitly reference criteria; *all* feedback is *generic*
Clear directions for improvement Feedback clearly marks what a writer has done well and is	**5:** *Consistently* gives clear directions for improvement + references specific student output in affirmation/glow; *offers specific examples for improved writing;* all feedback is clear, and some are ready to be implemented with minimal further research or study

specific about where and how they can improve Quality of feedback ranges from **vague → always offers clear directions**	**4:** *Mostly o*ffers usable feedback; *few directions* are *not spelled out* and suggestions are reliably tied to identified sections of the writing; most feedback is clearly usable, some is less clear
	3: *Even* mix of *specific* and *vague* suggestions; *some directions are not clearly spelled out* or suggestions are not tied to identified sections of the writing
	2: *Minimal* references to specific student writing **or** lacks actionable next steps/recommendations
	1: Does not reference specific student writing **and** does not give concrete steps aligned with rubric/prompt
Accurate Feedback must be accurate for cognitive and affective reasons Quality of feedback ranges from **inaccurate → accurate**	**5:** *All* feedback is accurate
	4: *Most* feedback is accurate; one piece of feedback is *somewhat inaccurate*
	3: *Some* feedback is accurate; 1+ pieces of feedback are **clearly inaccurate**
	2: Feedback is *mostly* inaccurate
	1: Feedback is *inaccurate*, irrelevant to student writing
Prioritization of essential features Feedback responds to the appropriate elements of writing given current writing and idealized writing in this genre. Feedback mentions the next, proximal step in their development.	**5:** *All* feedback focuses on the most **appropriate priority** to work on given the writing and essential features of SBAW in history
	4: *Most* feedback is attainable and **reasonable to work on next**; prioritizes *one of the most* appropriate things given current writing and essential features of SBAW in history

Feedback ranges in quality from **nonessential →essential and attainable** **Nonessential example:** (Writing lacks a clear claim) *Consider articulating and addressing a counterargument to improve the essay*	**3:** *Even mix* of prioritizing essential and nonessential features; *some* feedback is not attainable or appropriate; other feedback is not essential for the student given their writing
	2: *Most* feedback is *unattainable*, *nonessential*, or *inappropriate* given the student's current writing and the priorities of the rubric
	1: Feedback is too difficult for students given their writing level, or feedback does *not mention* any essential features
Supportive tone Quality of feedback ranges from *unsupportive and directive* → *affirming* and *supportive*	**5:** Feedback is *consistently* affirming; uses suggestive and respectful language; balance of positive comments/appraisals + suggestions for improvement
	4: *Most* feedback uses suggestive/supportive language, but some does not; it may have an imbalance of affirmations and suggestions for improvement
	3: *Even mix* of suggestive/directive language; may lack specific affirmation
	2: *Most* language is directive (not suggestive) or lacks any positive affirmation
	1: *No* positive comments and no suggestive, respectful language; the tone may be condescending, not polite

To ensure adequate reliability, a team of three was trained by the first Author in the use of the framework. The first author, one researcher with a PhD, and two undergraduate research assistants comprised the coding team. The team began by discussing the criteria for each level (1-5) in a specific component and practiced coding jointly. During coding, they refined criteria in the codebook and identified anchor texts that represented exemplars for the scoring categories of each feedback component.

Iterative cycles of coding, discussion, and refinement of the criteria and anchor texts continued until coders exhibited high degrees of interrater agreement. Then, raters double-coded 15% of the sample (n = 60) to ensure reliability before individually coding feedback. Exact agreement ranged from 68 to 87% across the five components (note that reliability was above 80% agreement for all categories but accuracy). Within 1-point agreement ranged from 97 to 100% (MacArthur et al., 2019; Troia et al., 2019). Cohen's Kappa ranged from .71 to .84. See Table 2 for the interrater agreement for the five feedback components.

Table 2

Interrater Agreement for Five Components of Feedback

Component	Within 1 point	Exact	Cohen's Kappa
Criteria-based	100.00%	83.87%	0.80
Clear directions	100.00%	87.10%	0.84
Accurate	96.77%	67.74%	0.71
Essential features	96.77%	80.65%	0.76
Supportive tone	100.00%	80.65%	0.76

Note. n = 60.

Analytic Approach

To answer our research questions, we calculated basic descriptive statistics for the quality of each of our coded types of feedback for ChatGPT and human evaluators. We then ran a one-way analysis of variance (ANOVA) model for each of the feedback characteristics to determine if the human feedback differed from the AI feedback (the independent variable) with respect to that component (each a dependent variable).

We addressed the second and third research questions by running ANOVAs (with Scheffe corrections as appropriate) using three categories of essay quality and the students' EL status. Well-trained human raters previously double-scored each essay for quality. We separated the essays into three groups: a high-scoring group with ratings more than a standard deviation above the mean, a low-scoring group with ratings more than a standard deviation below the mean, and an average-scoring group of the 66% around the mean (i.e., 17-83 percentiles). Data analysis was conducted using STATA (version 15).

Results

RQ #1: How Does the Quality of Formative Feedback Provided by ChatGPT Compare to the Quality of Feedback Provided by Human Evaluators?

Descriptive statistics in Table 3 show the scores for human and AI feedback in each category on a scale of 1 to 5 (n = 198). The partial eta-squared effect size describes the proportion of the variance in the dependent variable attributable to a particular independent variable, while Cohen's d effect size calculates the size of the difference between the means. We find that human feedback is better than AI feedback in every category except *criteria-based*, where AI outperformed human evaluators by .24 points on average (p = .03). Fitting a one-way ANOVA model, the differences between human and AI feedback ratings were statistically significant (*provided clear directions for improvement*, $p < .001$; *accurate*, $p < .001$; *prioritized essential features*, $p < .001$; *supportive tone*, $p < .001$).

Table 3

Quality of Human and AI Feedback by Category

Category	Effect size (Partial eta-squared)	Effect size (Cohen's *d*)	Human/ AI	Mean	SD	Skewness	Kurtosis

Criteria-based	.01	-0.22	Human	3.40	1.09	-.11	2.00
			AI	3.64	1.16	-.33	1.86
Clarity of directions for improvement	.05	0.41	Human	3.65	.89	-.21	2.51
			AI	3.27	.98	.11	2.30
Accurate	.07	0.61	Human	4.55	.74	-1.65	5.20
			AI	4.0	1.04	-89	3.17
Prioritization of essential features	.12	0.82	Human	3.92	.98	-.49	2.31
			AI	3.09	1.05	.11	2.42
Supportive tone	.03	0.32	Human	4.28	.84	-.87	2.83
			AI	4.02	.83	-.98	4.58

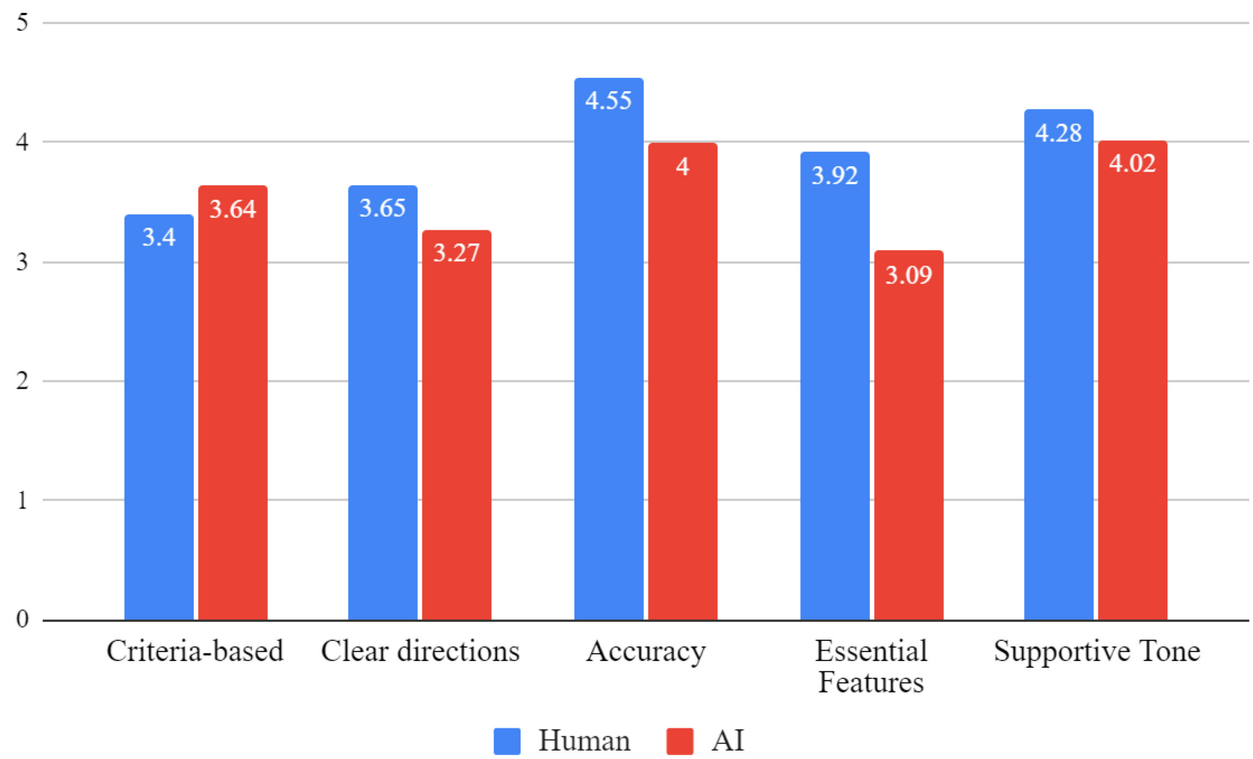
Note. Scores ranged from 1-5.

Mean differences between the feedback conditions ranged from -.24 (favoring AI) for criteria-based (3.64-3.40) to .83 (favoring human raters) for prioritization of essential features (3.92-3.09), indicating that human and AI feedback was within one point difference in all cases. Partial eta-squared effect sizes for variance explained by the difference in rater were all small, ranging from .01 (for criteria-based feedback) to .12 (for prioritization of essential features), indicating that the difference in raters explained very little of the differences in feedback ratings.

Although there were some small to moderate differences between human feedback and ChatGPT, the ChatGPT feedback was still of relatively high quality, with average quality ratings ranging from 3.09 to 4.02 across all dimensions (the average quality of human feedback ranged from 3.4 to 4.55 across dimensions). Figure 1 shows the average quality in key components of feedback by condition.

Figure 1

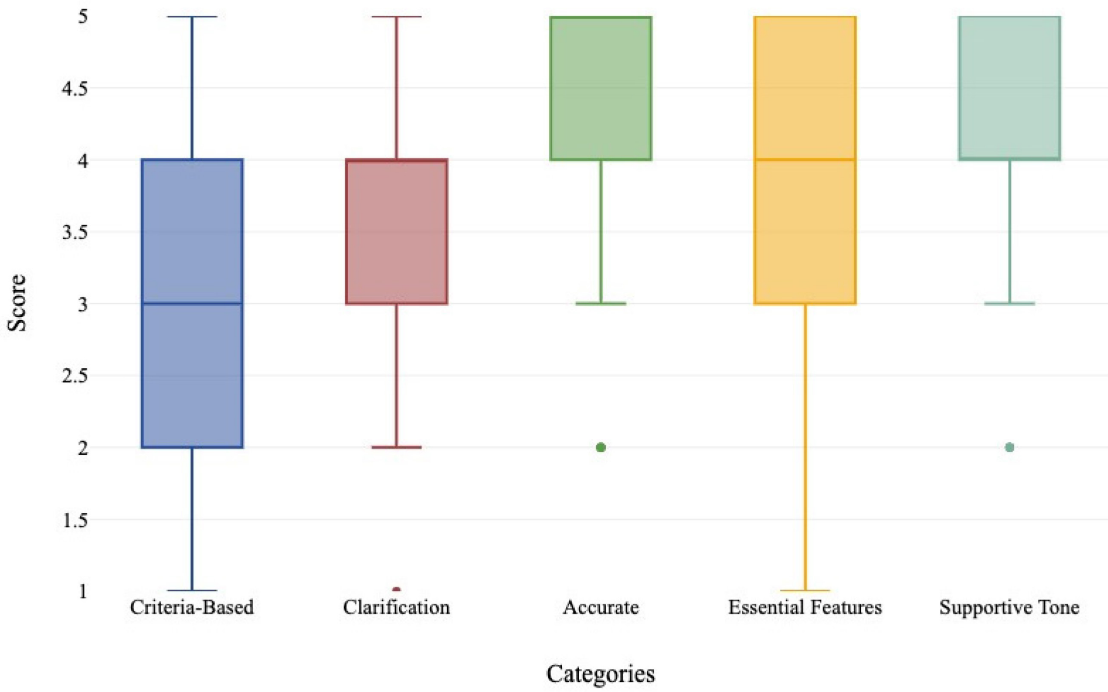
Average Scores for ChatGPT and Human Feedback



The following boxplots show the range of scores across the five criteria for human raters (Figure 2) and the AI rater (Figure 3). The middle boxes show the middle 50% of scores, and lines extend to the lower and higher 25% of scores, respectively. Dots represent outliers greater than 1.5 times the upper quartile.

Figure 2

Human Feedback Ratings by Criteria

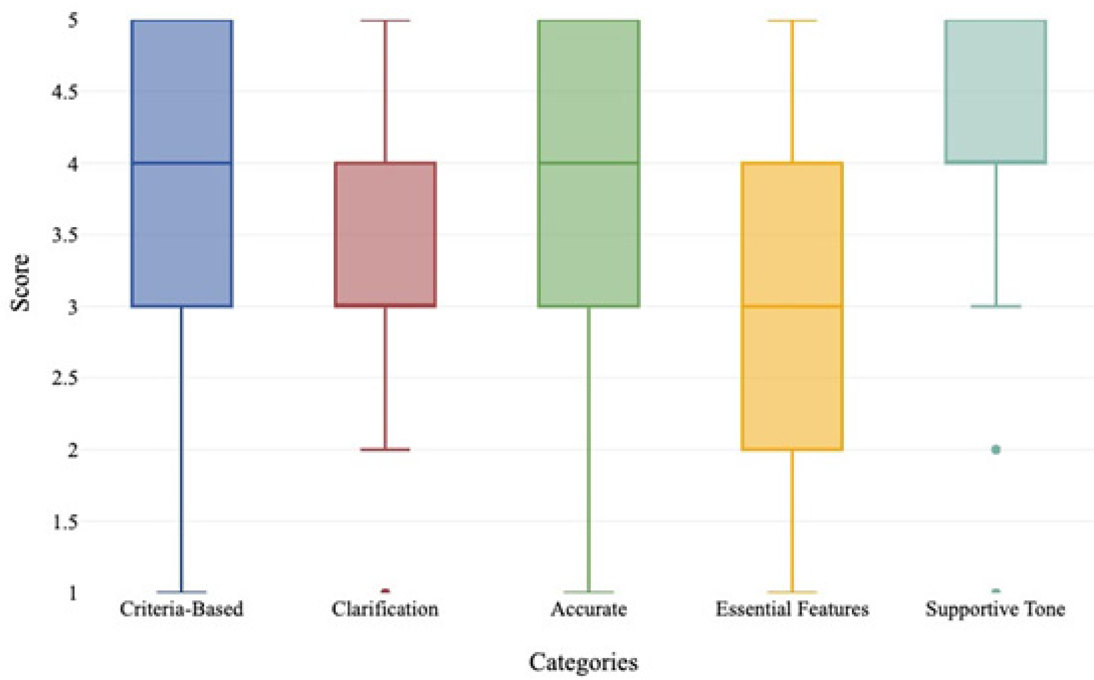


The box plot for human feedback reflects a normal distribution of scores for *criteria-based* around a mean of 3 and relatively higher scores for the other criteria, especially the accuracy of formative feedback and the use of a supportive tone. Though the majority of human scores for prioritization of essential features were high, the lower 25% ranges from 1 to 3.

The box plot for AI feedback reflects a normal distribution of scores for *prioritization of essential features* around a mean of 3.09, which shows the relatively lower quality of feedback for this component from the AI. The distribution of AI scores for criteria-based shows relatively higher scores for the middle 50% of scores, but the lower 25% of scores range from 1 to 3. Similar to human feedback, there were relatively higher scores for *accuracy* and *supportive tone*, though there were more outliers at the bottom end of the distribution for the AI was more likely offer inaccurate feedback. The distribution of scores for the *clear directions for improvement* dimension was similar to the human evaluators, showing similar performance in this component of feedback.

Figure 3

AI Feedback Ratings by Criteria



Note. The box plot reflects a normal distribution of scores for [prioritization of] essential features.

RQ #2: How Does the Quality of Formative Feedback Vary for High-, Average-, and Low-Scoring Essays?

The AI showed a statistically significant difference in ratings with respect to *accuracy* ($p < .001$), *prioritization of essential features* ($p = .03$), and *supportive tone* ($p = .001$) across different quality essays. Specifically, the accuracy of AI feedback declined as the quality of the underlying essay rose, with ratings of 4.42, 4.05, and 3.31 for low-, average-, and high-quality essays, respectively. *Prioritization of essential features* in AI feedback was best for average-scoring essays and worst for high-scoring ones, 2.98, 3.22, and 2.69 (low, average, high, respectively). *Supportive tone* in AI feedback was lowest for the low-scoring essays, suggesting a need to refine the AI prompting with suggested phrases or additional instructions to use a suggestive rather than a directive tone (3.56, 4.1, 4.16, low to high). Human raters showed a statistically significant difference in ratings with respect to *prioritization of essential features* (p = .002), with ratings of 4.44 for low-quality essays but less effective feedback for average- (3.79) and high-scoring (3.84) essays. Human feedback did not otherwise vary by scoring level of the underlying essays. The means for the various conditions are provided in Table 4. Figures 4a and 4b show the mean score for each component of feedback by the scoring level of the essays.

Table 4

Mean Levels of Criteria Scores for High-, Average-, and Low-Scoring Papers in Each Condition

	Quality: Low		Quality: Average		Quality: High	
	Human	AI	Human	AI	Human	AI
Criteria-based	3.19	3.56	3.36	3.63	3.78	3.81
Clarity of directions for improvement	3.81	3.08	3.60	3.33	3.69	3.25
Accurate	4.75	4.42	4.53	4.05	4.41	3.31
Prioritization of essential features	4.44	2.98	3.79	3.22	3.84	2.69
Supportive tone	4.33	3.56	4.31	4.10	4.09	4.16

Figure 4a

Mean Human Feedback Rating for Each Component of Feedback

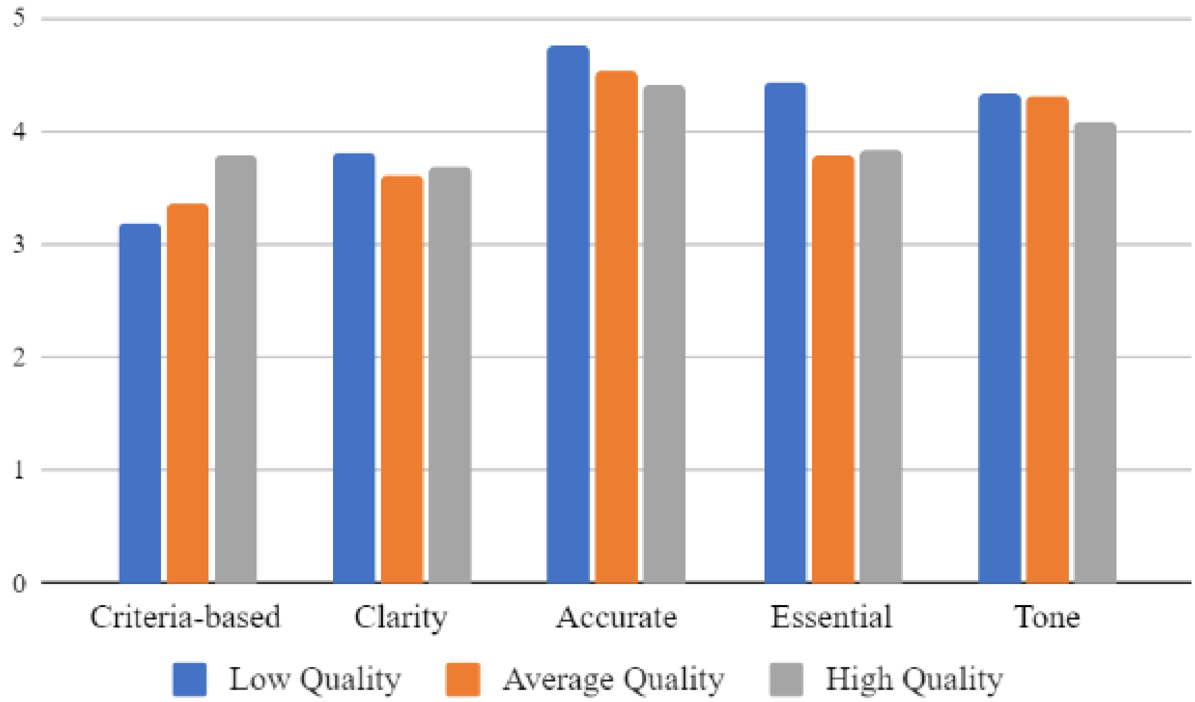
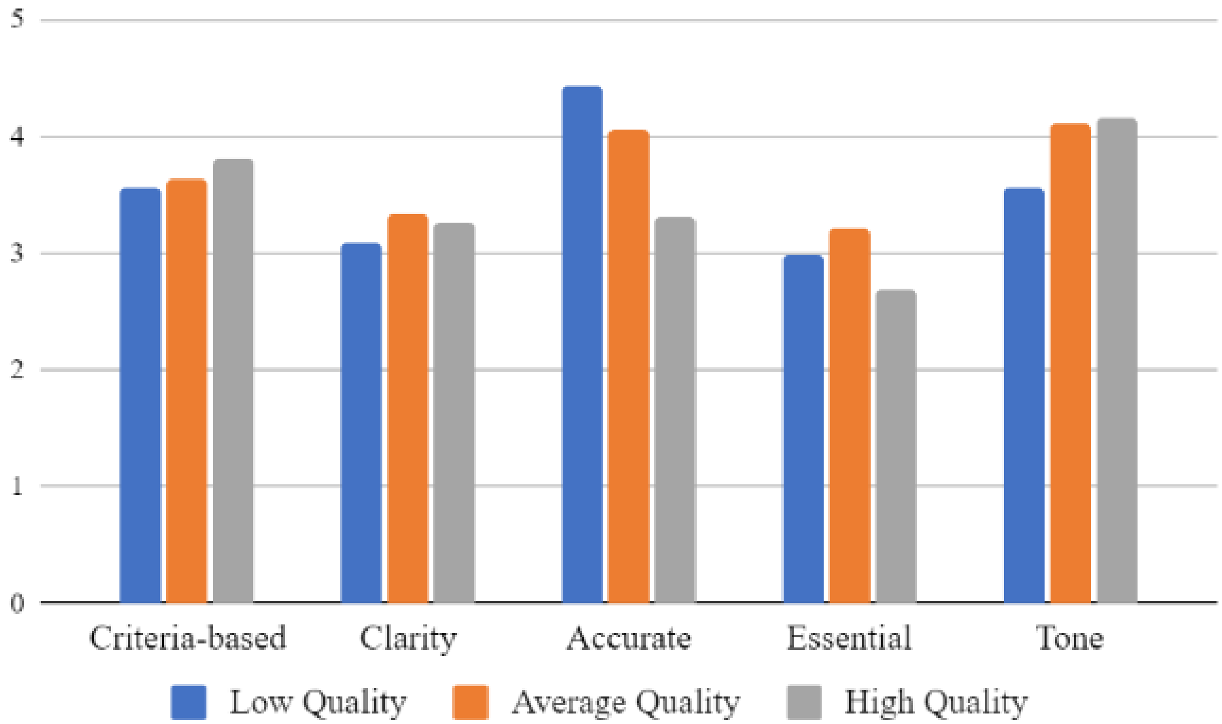


Figure 4b

Mean Rating by Human and AI Scorers for Average-Quality Essays



RQ #3: How Does the Quality of Formative Feedback Vary for Students with Different Language Statuses?

Feedback given to students who were either initially fluent or English-only speakers (collectively, EO/IFEP) was not statistically different compared to that given to students classified as English learners (ELs) and reclassified fluent speakers (RFEP) using Scheffe correction for multiple comparisons. This was true with respect to both the AI and the human raters.

Discussion

Differences between Human Evaluators and ChatGPT

We found that human raters, at least the well-trained, paid, and relatively time-rich evaluators in this sample, provided higher quality feedback in four of five critical areas: *clarity of directions for improvement*, *accuracy*, *prioritization of essential features*, and use of a *supportive* *tone*. The impressive skills of experienced and resourced human educators to provide quality formative feedback were notable. However, the most important takeaway of the study is not that expert humans performed better than ChatGPT—hardly a surprising finding—but rather that ChatGPT's feedback was relatively close to that of humans in quality *without requiring any training*. To our knowledge, no previous study has compared automated and human feedback on writing because the quality of automated feedback has been so poor that such a study would be futile. Presently, the small differences between the two modes of feedback suggest that feedback generated by ChatGPT can likely serve valuable instructional purposes, particularly in the early stages of writing to motivate revision work by students in a timely fashion.

Even considering the largest effect size between humans and ChatGPT for *prioritization of essential features*, the difference between a "3" and a "4" for feedback quality may be insignificant in some instructional contexts. ChatGPT-generated formative feedback may be helpful given the present demands of providing timely feedback and the ability of teachers to contextualize and frame the use of ChatGPT for their students in non-experimental settings. To better illustrate the differences between human evaluators and ChatGPT, consider the following feedback offered by a human evaluator and ChatGPT (Figure 5). Both pieces of feedback were scored similarly, with the human scoring 1 point higher for the degree to which the feedback was criteria-based, accurate, and used a supportive tone.

Figure 5

Sample Feedback Provided by a Human Evaluator and ChatGPT

<table>
<tr><td colspan="2">Student Essay: In the U.S. like any other cities during 1955, Montgomery, Alabama's busses were being segregated, the enforced separation of racial groups in a country, community, or establishment. In this case it was the black and whites being segregated in the bus; this meant 10 rows of the seats were only reserved for the whites; this had only left the African Americans to be forced in the rear of the bus having to be standing up or sitting down if they ever got the chance. In March 1955 Black leaders of Montgomery like Rosa Parks, Marthur Luther King Jr, and The women's political council met with the city officials to recommend and try to make changes for the Montgomery bus system. Why did the Montgomery Bus boycott Succeed? The Bus Boycott has succeeded because most of the people who rode Montgomery's buses were African Americans and this made the company lose a lot of business and revenue from the boycott so this left them with no easy solutions to fix.

Rosa Parks is also one of the many reasons why the bus boycott succeeded because she had refused to give up her seat for a white man on a bus and this led to her arrest. This injustice sparked the Montgomery Bus Boycott. ¨Parks' courageous act and the subsequent Montgomery Bus Boycott led to the integration of public transportation in Montgomery¨. Another reason the boycott was able to succeed was because of the strong leadership- King, MIA, NAACP, WPC, and role of churches. Black ministers from the church were responsible for calling their community members to join the boycott and the NAACP group sued the city in the U.S. district court. And the plan of action was announced on the front page of the newspaper.</td></tr>
<tr><td>Human Feedback: You did a wonderful job explaining what the Montgomery Bus Boycott was historically and you have a strong claim toward the end of your first paragraph: "The Bus Boycott has succeeded because most of the people who rode Montgomery's buses were African Americans and this made the company lose a lot of business and revenue from the boycott so this left them with no easy solutions to fix."

Next time, I would love to see some evidence from the sources to help back up your claim. This will show your readers that your claims are supported! Keep on practicing your writing and you'll keep improving.</td><td>ChatGPT Feedback: Great job on providing a brief background on the Montgomery Bus Boycott and highlighting the key players involved. You have also provided some reasons why the boycott succeeded, such as the loss of revenue for the bus company and the strong leadership of King, MIA, NAACP, WPC, and churches. To improve your essay, consider providing more specific examples to support your points. For instance, you could mention how much revenue the bus company lost during the boycott or provide specific actions taken by the leaders and organizations involved. Additionally, you could expand on the impact of the boycott beyond Montgomery and how it influenced the Civil Rights Movement as a whole.</td></tr>
</table>

Notably, both evaluators identify the need to integrate more evidence, which we see as appropriate for this student's writing. Next, the human uses criteria-based language, "evidence from the sources," while the AI uses more vague language, "examples." Still, the feedback is relatively similar and one was generated instantaneously and without training on a writing corpus. A teacher may find the ChatGPT feedback useful, given time constraints, even if the human feedback was better. Further, in an applied context and with sufficient AI literacy skills, one could ask ChatGPT, "What evidence could I integrate?" or "Can you give an example of how to integrate evidence into my essay?" to get additional, immediate, customized support.

An additional example of human and AI-generated formative feedback illustrates the downside of ChatGPT compared to human evaluation. In Figure 6, we see ChatGPT feedback that scored 1 point lower than human feedback on all categories except the prioritization of essential features. The 1-point magnitude of differences is greater than all of the average differences which ranged from .24 (favoring AI) to .83 (favoring human raters).

Figure 6

Sample Feedback Provided by a Human Evaluator and ChatGPT

<table>
<tr><td colspan="2">Student Essay: Why was the Montgomery Bus Boycott successful? Well the reason the boycott was successful was because Rosa Parks started to give flyers out to black people to stop using the buses. It did work because after a while the buses were not getting the money they used to get when the white and black people were there. Only the white were on the buses.

The second way they got the boycott to be successful is, Rosa Parks was in the front of the bus but black people were in the back but she sat on the front which they were not supposed to be on. She…</td></tr>
<tr><td>Human Feedback:Thank you for writing this essay on how the boycott succeeded! I think your claim, "Well the reason the boycott was successful was because Rosa Parks started to give flyers out to black people to stop using the buses" is interesting. But did Parks distribute the flyers? Better check that source again (I think it was Jo Ann Robinson and her students).

Next time, you might use this sentence starter when adding evidence to support your claim, According to the "Bus Boycott Timeline,"___________. This will help you support your claim with clear evidence</td><td>ChatGPT Feedback: Great start! You have identified some key factors that contributed to the success of the Montgomery Bus Boycott. However, your essay could benefit from more detail and explanation. For example, you mention that Rosa Parks gave out flyers, but you don't explain what was on the flyers or how they helped the boycott. Try to provide more specific examples and evidence to support your argument.

Additionally, it's important to note that Rosa Parks' refusal to give up her seat was not the</td></tr>
</table>

directly from the source. Your reader might also be convinced because you get your evidence from a source too. You are on the right track to becoming a great writer!	only reason for the boycott's success. Can you think of any other factors that contributed to the boycott's success? Keep up the good work!

First, the human evaluator correctly identifies the student's mistake in confusing Rosa Parks for Jo Ann Robinson. They also give clear directions for including more evidence to support the student's claim. Conversely, ChatGPT does not realize Rosa Parks has been mistaken for Jo Ann Robinson despite having been provided with all the source materials. This reminds us that ChatGPT does not actually "understand" the text it is given but instead relies on a predictive algorithm to generate the response. ChatGPT also contradicts itself, telling the student they identify "some key factors that contributed to the success of the Montgomery Bus Boycott," before later telling the student, "it's important to note that Rosa Parks' refusal to give up her seat was not the only reason for the boycott's success." This inaccurate comment fails to recognize something the student has done to meet the requirements of the prompt. Inaccurate feedback can discourage students, especially novice writers, from engaging in revision (Author, 2010; Moore & MacArthur, 2016).

Heterogeneity by Essay Quality and Language Status

While the main purpose of the study was to highlight differences between human evaluators and AI, findings related to the heterogeneity of feedback quality were notable. First, we noted no significant differences in the quality of feedback provided to English learners by either humans or the AI. We did, however, see differences in the quality of feedback given across high-, average-, and low-scoring essays.

When we compare the mean scores by essay quality level and type of feedback, we see most of the differences in ratings were one-third to one-half of one point in our five-point measure. Score differences by essay quality were more than a point difference in only three instances: prioritization of essential features (both high- and low-scoring papers) and accuracy (high-scoring papers). ChatGPT and human evaluators both struggled to provide good feedback related to the prioritization of essential features for higher-scoring essays, which is a notable finding for differentiated writing instruction. This could be ameliorated by better prompting of the AI tool. For example, we could have made it clear which components were priorities and in which order, specifically telling ChatGPT to note the next applicable priority. Better prompting will probably not improve the accuracy of the AI feedback for the high-scoring essays, but as the AI models improve generally it is possible that the accuracy would similarly improve.

Unlike the ChatGPT, which consistently did better on criteria-based feedback across the essay quality levels, the human feedback was almost as good on the high-quality papers as on the low-quality ones. The human feedback was better at the clarity of directions for improvement, with the ChatGPT being especially poor in the case of low-quality papers. Once again, refinement of the prompt could provide ChatGPT with specific examples of clear directions to improve this aspect of its feedback. ChatGPT struggled with accuracy when providing feedback to the high-quality papers and with maintaining a supportive tone to the low-quality papers, both of which could have a negative impact on these students' motivation to respond to feedback. Instructors using ChatGPT in classroom settings should be aware of this tendency and proactively address the potential for too much critical feedback by instructing students how to selectively interpret or ignore feedback not aligned with their writing goals.

Considering Utility Alongside Fallibility

While ChatGPT as a tool for writing seems promising, the shortcomings of AI tools (e.g., the inaccurate comment in Figure 6) need to be understood and addressed by educators who might use them to augment writing instruction in the classroom. This means developing students' and teachers' AI literacy is a key priority if generative AI tools are to be used productively and dynamically in writing instruction (Author, 2023). Prior research suggests that AI tools will best augment writing instruction if both instructors and students understand the mechanisms of AI-based writing evaluation systems and use them appropriately (Author, 2010; Wilson & MacArthur, in press).

For teachers, AI literacy involves reminding students that not all AI-generated feedback is *always* accurate and that as the author they have the ultimate say on how they want to express their thoughts. This contextualization of received feedback is also true of human-provided feedback. During the process of reconceptualization, students should evaluate feedback for accuracy and usefulness before integrating it into subsequent drafts. Teachers can provide time and scaffolding to support students in this reflection. As a next step, we propose that instructors combine AI-based feedback on earlier drafts with human feedback on a later or final draft. This keeps the human audience a priority and allows the instructor to correct any deficiency in the AI feedback, such as inaccurate comments or not prioritizing the essential next steps to work on, especially for higher-scoring essays.

It is also noteworthy that it took multiple attempts at prompt engineering to elicit the best feedback from ChatGPT. This iterative work reflects the challenges educators and students will face in learning to leverage generative AI tools in the classroom effectively. Presently, our prompt engineering was relatively brief, but prompting did impact the type of feedback that was generated. Therefore, training and support must be provided so teachers can develop AI literacy—understanding AI tools' underlying challenges and affordances and how to effectively use them for specific instructional purposes. Then, teachers can make critical pedagogical choices about using generative AI tools in the best ways for their specific instructional goals or student needs. AI literacy is one essential skill needed to use emerging technologies productively, ethically, and for various creative and social purposes (Author, 2023b). For example, practice using iterative cycles with AI to garner additional insights into one's writing represents a dialogic interaction that can benefit English learners' writing development (Su et al., 2023; Wagner et al., 2017).

Beyond the differences in feedback quality across the five feedback components, there are trade-offs to human-generated feedback that need to be considered—generating feedback is costly in terms of time and discourages teachers from assigning writing (Applebee & Langer, 2011). In the present study, evaluators had extensive experience in writing instruction, were trained, and had adequate time to provide feedback. Not all educational contexts have such resources.

Conversely, AI like ChatGPT can generate feedback instantaneously, responding to specific author instructions without training on a human-scored training set and as often as students or teachers request it. ChatGPT does not need to sleep, nor does it get tired of the same query. Timeliness matters as feedback is more effective when it is provided close in time to when the writing was completed (American Psychological Association, 2015; Clare et al., 2000; Ferster et al., 2012; Hattie & Timperley, 2007). While we gave ChatGPT only one opportunity to provide feedback and did so after the completion of writing, a student or teacher could ask for feedback during the drafting process. Users could also review feedback and ask for different, new, or more specific recommendations.

Finally, our study took place in the first year of ChatGPT's existence; with time and experience, not only will we be able to generate better instructions, but model performance will continue to improve. Given these important advantages, and despite the statistically significant outperformance of well-trained human evaluators in four out of five categories of formative feedback quality, ChatGPT and other generative AI may have a place in writing instruction if teachers and students were taught (1) how to use and effectively prompt the generative AI, and (2) that the feedback would not be infallible and so (as with any feedback, including peer feedback) the writer should *consider* but not necessarily *take up* all suggestions.

Implications for Future Studies

Further research is needed to understand what writers actually *do* with the feedback given by AI—does it cause them to do more revision than the little we typically see from students at this age and if so, does the revision improve the essay quality? While we used analytic coding to determine the quality of formative feedback, impact on student writing motivation, efficacy, or performance are better measures of feedback's effectiveness. We see the present study as justifying such a study in a classroom setting in the future.

Formative feedback is also a way for teachers to gain important knowledge about student writing development and to design evidence-based instruction (Author, 2016). Even if the present study shows feedback provided by AI as relatively high quality given the time it requires, we cannot measure the potentially negative effect of *not* engaging with early student drafts on teachers' knowledge of their students and subsequent evidence-based instructional planning. Using generative AI, instead of teachers, to frequently provide feedback would be especially problematic if teachers do not oversee the process, looking over students' shoulders at times to see how they are engaging with the AI feedback, asking students' perceptions of the feedback, and considering the local context for best use cases and limitations. The pedagogical implications for both students and teaching must be considered carefully.

Limitations

True “blinding” of feedback types was not achieved in the present study. Though we made efforts to present feedback similarly and to ensure the experience of evaluating feedback was similar across ChatGPT and human conditions, it was consistently obvious which feedback was composed by AI. It is possible that raters were biased against or for AI. Though we doubt this was the case, we are unable to say definitively if bias against human or AI scorers played a role in evaluation of feedback.

This study was done in the context of a larger study that included the same writing prompts across grade levels and it was expected that the feedback providers would adjust their feedback across grade levels based on the quality of student writing. In short, the independence of formative feedback quality and grade level was presumed. It might be beneficial and more valid in other contexts to give humans and ChatGPT context about the writer (e.g., grade level) so that they can better provide feedback to the students.

Additionally, a few factors may have “penalized” ChatGPT in the present study. First, ChatGPT feedback was longer on average, which may result in lower scores in our analytic framework as ChatGPT had more room to make errors. Sometimes, ChatGPT gave good feedback, interspersed with a lot of “fluff.” Though we typically scored such feedback lower, a student may plausibly ignore this extraneous feedback and benefit from other comments. We also used the less capable ChatGPT model, GPT-3.5, rather than the more capable GPT-4, as the former can be accessed at no cost and thus be more typically used by teachers and students today. However, conventional wisdom and our own small scale pilots suggest that GPT–4 is able to give higher quality feedback.

Second, the human evaluators in the present study were plausibly more experienced and more supplied with free time than average history teachers providing feedback to their students (Applebee & Langer, 2011; Author, 2009). In the larger intervention study, district history teachers reported no training in providing formative feedback to writing. The human feedback they received for their students, as part of the larger writing intervention, was immensely appreciated by these teachers, many of whom confessed they would not have had the time to provide it themselves. Meanwhile, the sample of human evaluators included writing teachers with decades of experience who received specific training and resources to provide feedback. They were also able to take multiple passes at reading the student essays and could revise their comments over three weeks.

Third, we did not iteratively prompt ChatGPT to refine its output despite its immense capacity to respond to user cues for additional or modified information. Additional research should examine how well ChatGPT does in modifying its feedback or accommodating additional requests as these are likely to occur in a classroom context.

Conclusion

Even if ChatGPT's feedback is not perfect, it can still facilitate writing instruction by engaging and motivating students and assisting teachers with managing large classes, thus providing them more time for individual feedback or differentiated writing instruction (Author, 2010). Given our results, we see a plausible use case for generative AI: providing feedback in the early phases of writing, where students seek immediate feedback on rough drafts. This would precede, not replace, teacher-provided formative or summative evaluation that is often more accurate and more tailored to student-specific characteristics, albeit less timely.

Because significant student revision prior to turning in a draft text is rare, we suspect that formative feedback from AI could motivate more revision than the current vacuum of formative feedback. It may also shorten the long delay between the first draft and subsequent revision as time-pressed secondary teachers wait for a long weekend to get to that "stack" of student papers. More on-demand, personalized feedback may be especially valuable to students whose first language is not English (Author, 2023c; Author, 2023d). Additionally, because ChatGPT did not require a training set like other AWE applications and was able to give feedback to writing in a specific-genre (argument writing in history), it appears to be applicable in a variety of genres and contexts, though future studies or educators should test these use cases.

We contend that AI's value can be realized by understanding both its strengths and limitations and deploying it in a way that maximizes the former and minimizes the latter. This includes educating teachers and learners on how AI functions and helping them use it critically and reflectively, incorporating more social forms of writing and assessment (Author, 2023b). Similar positive results for student learning have been found in research on other forms of AI for language development, including visual-syntactic text formatting (Author, 2019b) and conversational agents (Xu et al., 2022). We thus take our cues on how to approach large language models from this body of research, seeking out, as Authors (2010) expressed, "utility in a fallible tool."

References

American Psychological Association, Coalition for Psychology in Schools and Education. (2015). *Top 20 principles from psychology for preK–12 teaching and learning*. Retrieved from http://www.apa.org/ed/schools/cpse/top-twenty-principles.pdf

Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research, and Evaluation, 10*(1), 3.

Applebee, A. N., & Langer, J. A. (2011). "EJ" Extra: A snapshot of writing instruction in middle schools and high schools. *The English Journal*, *100*(6), 14-27.

Author. (2006).

Author. (2009).

Author. (2010).

Author. (2011).

Author. (2012).

Author. (2015).

Author. (2016).

Author. (2018a)

Author. (2018b)

Author. (2019).

Author. (2019b).

Author. (2023).

Author. (2023b).

Author. (2023c)

Author. (2023d)

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology, 37*, 67-81. https://doi.org/10.1080/01443410.2016.1223275

Beach, R., & Friedrich, T. (2006). *Response to writing.* In Handbook of writing research, Guilford, New York, pp. 222-234

Biber, D., Nekrasova, T., & Horn, B. (2011). The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis. *ETS Research Report Series*, *2011*(1), i-99.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*, 5–31. https://doi.org/10.1007/s11092-008-9068-5

Breakstone, J., Smith, M., & Wineburg, S. (2013). Beyond the bubble in history/social studies assessments. *Phi Delta Kappan*, *94*(5), 53-57.

Chen, D., Hebert, M., & Wilson, J. (2022). Examining human and automated ratings of elementary students' writing quality: A multivariate generalizability theory application. *American Educational Research Journal, 59*(6), 1122-1156.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and instruction, 20*(4), 328-338.

Clare, L. C., Valdes, R., & Patthey-Chavez, G. G. (2000). *Learning to write in urban elementary and middle schools: An investigation of teachers' written feedback on student compositions* (No. 526). Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Common Core State Standards Initiative. (2012). *Implementing the Common Core State Standards.* Retrieved from Common Core State Standards Initiative: Preparing America's Students for College and Career: http://www.corestandards.org

Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M., Greenleaf, C., Lee, C.D., & Project READi. (2016). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist*, *51*(2), 219-246.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Hillocks, G. (1986). *Research on written composition: New directions for teaching*. ERIC Clearinghouse on Reading and Communication Skills.

Kluger, A. N., & DeNisi, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword. *Current Directions in Psychological Science, 7*(3), 67-72.

Lawrence, J. F., Galloway, E. P., Yim, S., & Lin, A. (2013). Learning to write in middle school? Insights into adolescent writers' instructional experiences across content areas. *Journal of Adolescent & Adult Literacy, 57*(2), 151-161.

Levine, S. (2014). Making interpretation visible with an affect-based strategy. *Reading Research Quarterly*, *49*(3), 283-303.

MacArthur, C. A. (2016). Instruction in evaluation and revision. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 272-287). Guilford.

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction?. *Reading and Writing, 32*, 1553-1574.

Moore, N. S., & MacArthur, C. A. (2016). Student use of automated essay evaluation technology during revision. *Journal of Writing Research, 8*(1), 149-175.

Motz, B., Canning, E., Green, D., Mallon, M., & Quick, J. (2021). The influence of automated praise on behavior and performance. *Technology, Mind, and Behavior, 2*(3), 1-12.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational research review, 9*, 129-144.

Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing, 15*(2), 68-85.

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning, 31*(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994

Roscoe, R. D., Varner, L. K., Crossley, S. A., & McNamara, D. S. (2013). Developing pedagogically-guided algorithms for intelligent writing feedback. *International Journal of Learning Technology*, *8*(4), 362-381.

Roscoe, R. D., Wilson, J., Johnson, A. C., & Mayra, C. R. (2017). Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior, 70,* 207-221.

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., & Wilson, J. (Eds.) (in press). *Routledge international handbook of automated essay evaluation.* Routledge.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65. http://dx.doi.org/10.1016/j.asw.2013.11.007

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*, 100752.

Troia, G. A., Shen, M., & Brandon, D. L. (2019). Multidimensional levels of language writing measures in grades four to six. *Written Communication*, *36*(2), 231-266.

Underwood, J. S., & Tregidgo, A. P. (2006). Improving student writing through effective feedback: Best practices and recommendations. *Journal of Teaching Writing*, *22*(2), 73-98.

Van Steendam, E., Rijlaarsdam, G., Sercu, L., & Van den Bergh, H. (2010). The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. *Learning and Instruction, 20*(4), 316-327.

Wagner, C. J., Parra, M. O., & Proctor, C. P. (2017). The interplay between student-led discussions and argumentative writing. *TESOL Quarterly, 51*(2), 438-449.

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.

Wilson, J., & MacArthur, C. (in press). Exploring the role of automated writing evaluation as a formative assessment tool supporting self-regulated learning and writing. In *Routledge international handbook of automated essay evaluation*.

Wilson, J., Myers, M. C., & Potter, A. (2022). Investigating the promise of automated writing evaluation for supporting formative writing assessment at scale. *Assessment in Education: Principles, Policy & Practice*, 1–17. https://doi.org/10.1080/0969594X.2022.2025762

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research, 58*, 87-125. https://doi.org/10.1177%2F0735633119830764

Xu, Y., Aubele, J., Vigil, V., Bustamante, A. S., Kim, Y. S., & Warschauer, M. (2022). Dialogue with a conversational agent promotes children's story comprehension via enhancing engagement. *Child Developmen*t, *93*(2), e149-e167.

Zhai, N., & Ma, X. (2022). Automated writing evaluation (AWE) feedback: A systematic investigation of college students' acceptance. *Computer Assisted Language Learning, 35*(9), 2817-2842.

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education, 143*, 103668.