

Research Article

Testing the Ability of Teachers and Students to Differentiate between Essays Generated by ChatGPT and High School Students

Tal Waltzer ¹, Riley L. Cox ^{1,2} and Gail D. Heyman ¹

¹Department of Psychology, University of California San Diego, La Jolla CA 92093, USA

²Cathedral Catholic High School, San Diego CA 92130, USA

Correspondence should be addressed to Gail D. Heyman; gheyman@ucsd.edu

Received 3 April 2023; Revised 23 April 2023; Accepted 29 April 2023; Published 26 June 2023

Academic Editor: Zheng Yan

Copyright © 2023 Tal Waltzer et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The release of ChatGPT in late 2022 prompted widespread concern about the implications of artificial intelligence for academic integrity, but thus far there has been little direct empirical evidence to inform this debate. Participants (69 high school teachers, 140 high school students, total $N = 209$) took an AI Identification Test in which they read pairs of essays—one written by a high school student and the other by ChatGPT—and guessed which was generated by the chatbot. Accuracy was only 70% for teachers, and it was slightly worse for students (62%). Self-reported confidence did not predict accuracy, nor did experience with ChatGPT or subject-matter expertise. Well-written student essays were especially hard to differentiate from the ChatGPT texts. In another set of measures, students reported greater optimism than their teachers did about the future role of ChatGPT in education. Students expressed disapproval of submitting ChatGPT-generated essays as one's own but rated this and other possible academic integrity violations involving ChatGPT less negatively than teachers did. These results form an empirical basis for further work on the relationship between AI and academic integrity.

1. Introduction

Artificial intelligence (AI) has the potential to radically transform the way people live, so it is unsurprising that there has been extensive speculation about its implications. For example, scholars have considered what it means for the workplace, such as how people can collaborate with AI to get work done [1] and the many ways in which AI could pose a threat to job security [2–4]. Scholars have also raised questions about the role of AI in medicine, for example in relation to medical diagnoses and health care systems [5, 6], and how it might affect people's awareness of public health issues [7]. Another far-reaching set of questions concerns its effect on education, including its potential to promote critical thinking [8] or identify which teaching approaches work best in different contexts [9], as well as the possibility that it will exacerbate social problems such as discrimination in education [10, 11].

Another potential concern about AI in education is that it may undermine academic integrity, and this has become particularly salient since the public release of ChatGPT in November 2022. ChatGPT is an artificial intelligence (AI) chatbot built on a generative large-language model. It has been trained on a wide range of texts and uses statistical modeling to predict the words or phrases that are most likely to appear next, which allows it to generate sophisticated responses to prompts on a wide range of topics [12, 13]. Notably, ChatGPT has natural language processing capacities that go far beyond what its publicly available predecessors were capable of [14]. Because ChatGPT can generate responses to a broad range of prompts that seem difficult to distinguish from human-generated writing, students may be tempted to use it to cheat on their schoolwork [14–18], and some schools have already banned it for this reason [19].

Despite widespread concerns about the negative implications of ChatGPT and other forms of AI for academic integrity, there has been extremely little empirical data to inform students, teachers, and policymakers about this issue. One important gap in the literature concerns the need to assess the extent to which there is an actual threat to academic integrity. Specifically, can teachers distinguish writing samples generated by ChatGPT from the writing of their students, and are there any factors that can predict this ability?

A second important gap in the literature concerns people's attitudes about ChatGPT in relation to cheating, as well as more general attitudes about the use of AI in education [20, 21]. Such attitudes are likely to affect how useful a role AI ends up playing in the classroom [22]. More broadly, attitudes about AI have implications for how the technology will be accepted and implemented [23]. In research on everyday ethical behavior, theories about moral attitudes (e.g., the Theory of Planned Behavior) point to ways in which people's attitudes regarding moral behaviors predict their tendency to engage in them [24–26].

In the present work, we seek to address these gaps in the literature by assessing teachers and students in a high school context. This period is of interest because it is a formative time for decisions relating to future academic and career paths [27, 28].

2. Literature Review

Even before the release of ChatGPT, researchers were starting to ask questions about how people's writing compares to writing generated by AI, and several studies have investigated precursors to ChatGPT (e.g., GPT-2, [29, 30]). One such study was conducted with nine participants who had a background in English literature [29]. In the first phase of the study, participants were provided lines from unfamiliar poems and short stories and asked to create continuations of them, either with or without assistance from GPT-2. In the next phase, participants saw unfamiliar continuations and guessed whether they were made with the assistance of AI. Participants performed well above chance but were far from perfect: they mistakenly thought AI-assisted texts were human-generated in 18% of cases and that human-written texts were made with AI assistance in 13% of cases. In a second set of studies, participants performed poorly on a task in which they were asked to distinguish between poems generated by GPT-2 and human-written poems [30]. For example, in one of these studies, the researchers found that performance was at chance (54%) when they hand-picked the best GPT-2 poems from a set that it generated, and performance was slightly better (66%) when the GPT-2 poems were chosen at random. These findings suggest that even with AI technology that is substantially less advanced than ChatGPT, detecting AI-generated writing is not always easy.

Since the release of ChatGPT, scholars have begun to investigate its ability to respond to academic tasks. Most of this scholarship has focused on the academic work required of undergraduate and postgraduate students. Yeadon et al. [31] used ChatGPT to generate five 300-word essays in

response to prompts from a physics module for second-year undergraduate physics students. For example, one prompt was "Does Kuhn or Popper give a more accurate description of physics?" Graders gave the ChatGPT essay an average score of 71%, which was comparable to the average score for undergraduates, and the ChatGPT essays received scores from Grammarly (2%) and Turnitin (7%) that indicated plagiarism was unlikely. Furthermore, Cotton et al. [15] found that ChatGPT performed at a level expected of third-year medical students on U.S. medical licensing exams, and Terwiesch [32] found that it performed at a B- to B-level on a final exam for an MBA course in operations management. ChatGPT Plus, which is based on GPT-4 and was released in March 2023, was able to earn a passing score on a bar exam and some other standardized tests (e.g., GRE Verbal, AP Biology, AP Environmental Science) [33].

There has been less research on the use of ChatGPT in pre-college educational settings. In one such study, de Winter [34] used it to generate responses to the multiple choice and short-answer questions on a national high school English exam in the Netherlands. ChatGPT performed about as well as the average student, and it was better at some kinds of questions (e.g., verbal analogies) than others (e.g., conditional reasoning). This finding raises the possibility that the writing of high school students may be hard to distinguish from writing generated by ChatGPT, which is one focus of the present research.

It is possible that there will prove to be individual differences in the ability to differentiate between high school students' writing and writing generated by ChatGPT. One potential predictor is meta-awareness as indexed by self-reported confidence in this ability [35]. Prior research suggests that the relationship between one's confidence and one's performance is highly context-dependent and can vary widely even for similar types of judgments [36]. For example, when an eyewitness identifies a suspect in a police lineup, confidence and performance are closely related when performance is measured for the first time on an unbiased test, but this relationship declines sharply when performance is measured on subsequent occasions or when the test is biased [37]. Confidence in one's ability to determine whether a piece of writing has been generated by a chatbot is important to understand in an academic context because overconfident teachers might falsely accuse students of cheating, which would be unfair, hard to defend against, and potentially damaging to students' future prospects.

Participants' relevant experience may also matter. One type of experience that may be relevant is prior use of ChatGPT. Such experience may help people recognize its writing style. For example, one pilot subject explained that she was able to differentiate between the essays based on her observation that "texts written by chatbots generally have longer and more complex sentences and have a conclusion sentence at the end." Experience with ChatGPT might also help people notice patterns across different responses. Because ChatGPT tends to give similar answers to the same prompt [15], it is possible that recognizing this tendency could boost identification accuracy on tests in which evaluators see multiple ChatGPT responses to the same prompt.

Teaching experience may matter as well. For example, expertise in grading the relevant types of writing might help teachers recognize typical features of student writing that others tend not to notice.

An additional question is whether the ability to differentiate between students' writing and writing generated by ChatGPT might vary as a function of the quality of the students' writing. We predicted that low-quality essays would be more distinguishable from writing generated by ChatGPT, based on the assumption that participants would not expect ChatGPT to generate poor-quality essays.

We addressed these questions by testing the performance of high school teachers and students on an *AI Identification Test* that assessed their ability to distinguish between short English essays written by ChatGPT versus high school students. We examined possible predictors of the ability to make this distinction, such as prior experience with ChatGPT. Participants also took an *AI Attitude Assessment* to measure their broader attitudes about the use of ChatGPT in academic settings, including their evaluations of its impacts on education and what uses should be considered acceptable.

3. Method

3.1. Overview. This study consisted of two phases. First, there was a *stimulus preparation phase* in which the stimuli were created. Second, there was a *testing phase* in which teachers and students took an AI Identification Test. The goal of this phase was to assess the overall accuracy of each group and identify any factors that might be correlated with their success on the test. Building on recent methods developed prior to the release of ChatGPT, we asked participants to read AI-generated and human-generated texts and infer their origin [29, 30, 38]. Specifically, we presented pairs of essays side by side and asked which one seemed to be generated by AI. We adapted the task for a high school context by testing high school teachers and students and by obtaining samples of students' writing that they had written for credit in an English class.

During the testing phase, participants also responded to an AI Attitude Assessment and self-reported on several possible correlates of performance on the AI Identification Test (e.g., subject matter expertise).

3.2. Participants. For the testing phase of the project, we recruited 69 high school teachers (23 of them taught English and 45 taught other subjects such as math, science, and history). Most of the teachers were women (54%), white (71%), and native English speakers (97%). We also recruited 140 high school students ($M_{\text{age}} = 16.86$ years, 44% girls, 57% white, 96% native English speakers). All data collection took place between February 27, 2023, and March 6, 2023, and followed our preregistration plan (<https://aspredicted.org/ec6ts.pdf>).

3.3. Procedure

3.3.1. Stimulus Preparation Phase. The stimulus preparation phase began by developing the prompts for the essays in collaboration with a high school English teacher. The goal was

to create prompts that were similar to those that students typically see in their English classes and that could reasonably be answered in a paragraph that is four to six sentences long. We wanted to keep the essays short so that each participant would be able to evaluate several pairs of them without losing their focus. The final prompts were as follows.

- (1) *Literature essay:* Why does literature matter? Please write this in third person. Your complete answer should be 4-6 sentences long.
- (2) *Proverb essay:* There are many proverbial phrases in our world. A few examples are "When life gives you lemons, make lemonade" or "the early bird catches the worm." Pick a proverbial phrase of your own and analyze its meaning. Please write this in third person. Your complete answer should be 4-6 sentences long.

The English teacher then assigned both essays to all of the students in four of his English classes (three sophomore classes and one senior class) between February 2, 2023, and February 9, 2023, and 97 essays were generated for each prompt. Students were told that their essays would be graded. To ensure that the essays would be written without any assistance from AI, students wrote them by hand without the use of electronic devices during proctored in-person sessions. Research assistants then transcribed all of the handwritten responses and removed responses from the pool based on the following predetermined criteria.

- (1) Essays must be written in the third person, as specified in the instructions.
- (2) Essays must fit the length requirement that was specified in the instructions.
- (3) Essays must be legible.
- (4) Essays must include content that is relevant to the prompt.

After these exclusions were applied, we were left with 40 literature essays and 42 proverb essays (most essays were excluded for not using the third person). We then used a randomization program (<http://randomresult.com>) to choose a pool of 25 essays from each set. After these essays were selected, a word processor was used to correct technical errors (e.g., singular/plural agreement, spelling, capitalization, and punctuation) so they would not serve as obvious cues and because many students use this type of software before submitting written assignments.

Two teachers graded the final essays in the same manner they would grade their own students' work. We averaged their grades together to assign a quality score to each student-written essay ($mean: 85.95$, $SD = 6.21$, range: 74.50 to 97.00).

The ChatGPT stimuli were prepared by entering the same prompts into ChatGPT (<https://chat.openai.com/>) on February 2, 2023, and regenerating the responses until there were 25 different essays for each prompt.

3.3.2. Testing Phase. In the participant testing phase, teachers and students served as participants. The student data were collected from ten different high school classes at the same school. No students who had contributed essays to the stimulus preparation phase were included in this phase.

All participants began by reporting the name of their school and their role as either a teacher or a student. After they indicated their familiarity with ChatGPT (“Please rate how much experience you have with using ChatGPT”) and their confidence in whether they would be able to distinguish between writing generated by ChatGPT and by high school students, they took the AI Identification Test.

The AI Identification Test consisted of six pairs of essays. There were three pairs shown for each of the two prompts, with each pair containing one essay generated by ChatGPT and one generated by a high school student, with each essay drawn at random without replacement from the appropriate pool of 25 essays. The order of the prompts was counterbalanced, with either the three literature pairs or the three proverb pairs appearing first. Examples are shown in Figures 1 and 2.

Participants were asked to read each pair of essays and select the one that they thought was generated by ChatGPT and were told that they would find out how well they had done when the task was over. After each guess, they were asked, “How confident are you about this answer?” (Slider from 0 = not at all confident to 100 = extremely confident). After making all of their guesses, they estimated how many of the pairs they had guessed right, from 0 to 6.

Next, participants took the AI Attitude Assessment, in which they predicted how many students would use ChatGPT to cheat, shared their concerns and hopes about its implications for education, and evaluated different possible uses of ChatGPT (see Table 1 for the prompts). The attitude assessment builds on previous methods that have measured people’s evaluations of everyday transgressions (e.g., academic cheating) [39, 40].

All participants were asked to rate the extent to which they already knew the subject matter (i.e., English literature). Teachers were also asked whether they had taught an English class and to describe their teaching experience. At the end, participants reported basic demographic information (e.g., age, gender). All prompts are available at <https://osf.io/8qv6z/>.

3.4. Data Analysis. We descriptively summarized variables of interest (e.g., overall accuracy on the AI Identification Test). We used inferential tests to predict AI Identification Test accuracy based on group (teacher or student), confidence, subject expertise, and familiarity with ChatGPT. We also predicted responses to the AI Attitude Assessment as a function of group (teacher or student).

Key hypotheses were tested using Welch’s two-sample *t*-tests for group comparisons, linear regression models with *F*-tests for other predictors of accuracy, and generalized linear mixed models (GLMMs [41]) with likelihood ratio tests for within-subject trial-by-trial analyses. GLMMs used random intercepts for participants and predicted trial performance (correct or incorrect) using trial confidence and essay quality as fixed effects.

4. Results

4.1. AI Identification Test

4.1.1. Overall Performance. Teachers correctly identified the essay that was generated by ChatGPT 70% of the time, and students correctly identified it 62% of the time. Binomial tests revealed that both groups performed above chance (chance: 50%; teachers: $p < .001$, 95% CI [65%, 74%], students: $p < .001$, 95% CI [58%, 65%]). Still, they were far from perfect: Most of the teachers (84%) and students (87%) got at least one trial wrong.

Teachers performed better than students, as confirmed by Welch’s two-sample *t*-test: $t(160) = 2.50$, $p = .013$. General confidence before the task did not predict performance, $F(1) = 0.09$, $p = .760$. Even after completing the task, participants’ estimation of how many they got right did not predict their accuracy, $F(1) = 0.95$, $p = .330$. Prior experience with the subject (English literature) did not significantly predict performance for either group, $F_s > 2.76$, $p_s > .098$, and English teachers performed no better than non-English teachers, $t(43) = 0.53$, $p = .599$. Prior experience with using ChatGPT also did not predict performance, $F_s < 3.48$, $p_s > .064$.

4.1.2. Trial-by-Trial Performance. Participants’ confidence on each trial did not significantly predict their performance, $D(1) = 0.48$, $p = .490$. However, the quality of the student’s essay did predict performance: The trials that participants answered correctly had student-written essays that had been rated as lower in quality ($m = 85.47$) during the stimulus preparation phase than the trials participants answered incorrectly, ($m = 86.72$), $D(1) = 12.03$, $p < .001$.

4.2. AI Attitude Assessment

4.2.1. Concerns and Hopes about ChatGPT. Participants estimated that about two-thirds of high school students would use ChatGPT to cheat if it were freely available (teachers: 67%, students: 64%). These estimates did not significantly differ by group, $t(147) = 1.15$, $p = .251$. On a scale from 0 (not at all concerned) to 100 (extremely concerned), both teachers ($m = 70.25$) and students ($m = 51.62$) expressed considerable concern about ChatGPT having negative effects on education; however, teachers were significantly more concerned, $t(181) = 5.18$, $p < .001$. Students expressed more optimism about ChatGPT having positive benefits for education ($m = 53.13$) than teachers did ($m = 38.91$), $t(166) = 4.02$, $p < .001$.

4.2.2. Evaluations of Different Uses of ChatGPT. Participants’ evaluations of how good or bad it would be for students to use ChatGPT varied depending on how it would be used (Figure 3). For example, on a scale from -10 (really bad) to +10 (really good), teachers ($m = -8.38$) and students ($m = -5.87$) rated it as pretty bad for a student to use ChatGPT to write an essay and then submit it. On the other hand, teachers rated it as a little bad ($m = -2.35$) and students rated it as a little good ($m = 1.58$) for a student to write an essay, ask ChatGPT to improve the

Both essays were written in response to the following prompt:

Why does literature matter? Please write this in third person. Your complete answer should be 4-6 sentences long.

Please click on the text that you think was written by the chatbot.

Literature is considered important because it offers a means of examining the human experience and exploring various perspectives, cultures, and beliefs. Through reading and studying literature, individuals are able to gain insight into the emotions, motivations, and experiences of others, as well as gain a deeper understanding of themselves. Additionally, literature provides a way to preserve cultural heritage and pass on important values and traditions from one generation to the next. Furthermore, literature can inspire, challenge, and entertain its readers, making it a valuable form of art and expression.

Literature matters because it shapes people's minds. Growing up with a good understanding of different pieces of literature can build intelligence and prepare people for adulthood. Literature can expand creativity, help with speaking in conversation, and just overall enlighten the human brain on different ideas in the world. Even going into adulthood, it is still important to experience the joy of reading and overall is a key part to a person's ability to be successful in life.

FIGURE 1: Example essays for the literature prompt, with the ChatGPT essay appearing first.

Both essays were written in response to the following prompt:

There are many proverbial phrases in our world. A few examples are "When life gives you lemons, make lemonade" or "the early bird catches the worm." Pick a proverbial phrase of your own and analyze its meaning. Please write this in third person. Your complete answer should be 4-6 sentences long.

Please click on the text that you think was written by the chatbot.

The proverbial phrase, "the early bird gets the worm" is commonly used and very well known. The word "early" is generally referred to in the time of the day, typically sunrise. The words "bird" and "worm" indicate the bird trying to catch its prey, or something working towards an intended goal. Overall, the phrase suggests that if a bird waits too long to eat, there will not be any food left. This indicates that if one does not act early, they will not be able to complete the goal desired.

One proverbial phrase is "don't count your chickens before they hatch." This phrase means to not assume that something will happen until it has actually happened, and not to make plans based on unfulfilled promises. It is often used as a reminder to be cautious and not to be overly optimistic about an uncertain outcome. The phrase suggests that it is important to be realistic and not to assume that everything will work out, as things may not turn out as expected. Overall, the phrase advises to avoid making assumptions or plans based on uncertain future events, and to wait for things to happen before making any conclusions.

FIGURE 2: Example essays for the proverb prompt, with the ChatGPT essay appearing last.

essay's vocabulary or structure, and then submit it. Teachers rated almost all of the hypothetical uses of ChatGPT more negatively than students did, $t_s > 3.42$, $p_s < .001$; the exception was using ChatGPT to generate practice problems while studying for a class (which both groups viewed positively), $t(166) = 0.64$, $p = .525$.

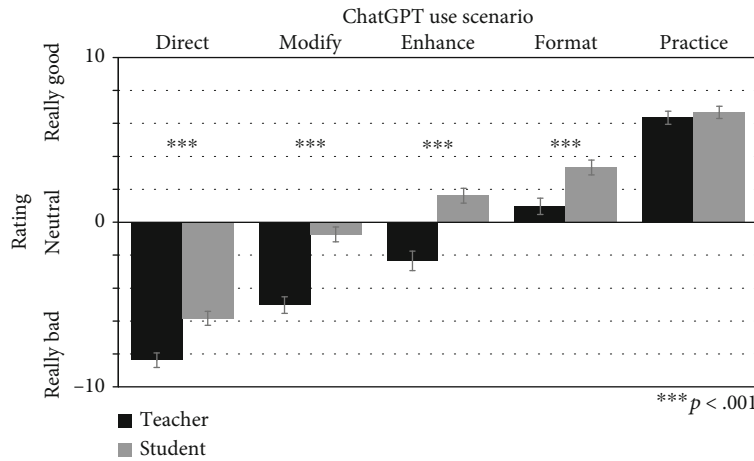
5. Discussion

ChatGPT has capabilities that go far beyond its publicly available predecessors. Not surprisingly, laypeople and researchers alike have been raising questions about its implications for individuals and the social institutions that shape

TABLE 1: Summary of AI Attitude Assessment Prompts.

Measure	Prompt	Response options
Estimated cheating	Assuming ChatGPT is freely available, what percent of high school students in the United States would ask ChatGPT to write an essay for them and submit it?	0% to 100%
Concern	How concerned are you about ChatGPT having negative effects on education?	0 = not at all concerned to 100 = extremely concerned
Optimism	How optimistic are you about ChatGPT having positive benefits for education?	0 = not at all optimistic to 100 = extremely optimistic
Comments	(Optional) Do you have any comments? You are welcome to explain your thoughts about any of the above questions. Please rate how good or bad the following actions would be, in your personal opinion.	An open-ended text box
Evaluation of uses	(e.g., a student uses ChatGPT to write an essay for them and submits the directly generated answer)	-10 = really bad to 0 = neutral to +10 = really good

Note: quantitative responses were made using a slider.



Direct: A student uses Chat GPT to write an essay for them and submits the direct generated answer.
 Modify: A student uses Chat GPT to write an essay for their class, then the student edits the output and submits the revised essay.
 Enhance: A student writes an essay, asks Chat GPT to improve the essay's vocabulary or structure, and then submits the exact revision.
 Format: A student writes an essay, then uses Chat GPT to change it into a specific format (e.g., MLA, APA) and submits the output.
 Practice: A student uses Chat GPT to generate practice problems while studying for class.

FIGURE 3: Participant ratings of various ChatGPT uses, grouped by teachers versus students. Bars represent standard error of the mean.

their lives. Many of these questions revolve around academic integrity in education [14, 42]. The present research takes a first step toward examining this issue in a high school setting.

We developed an AI Identification Test to assess the ability of teachers and students to distinguish between essays written by high school students versus ChatGPT. On average, the teachers (70% accuracy) scored slightly better than the students (62% accuracy). Both groups scored substantially better than chance (50%), but the overall level of performance indicates that both groups found the AI Identification Test quite difficult.

A major focus of the present research was to identify factors that are associated with teachers' success on the AI Identification Test, given that teachers are typically responsible for evaluating students' essays. Prior research has found that the extent to which confidence serves as a good indicator of one's performance is highly dependent on the context [30, 37]. We found that in the present context, confidence

was such a poor indicator of accuracy that it was essentially useless. We also asked the teachers about their prior experience with ChatGPT, and whether they had taught English before, and neither factor predicted their accuracy.

As part of the stimulus preparation phase, the essays written by high school students were rated by a pair of teachers. The essays that were judged to be higher in quality were more difficult for participants to distinguish from the writing samples generated by ChatGPT, which suggests that ChatGPT tends to generate output that is seen as higher in quality than the writing of many high school students. One teacher who scored above average (83%) seemed to pick up on this difference, noting, "I tended to think the better written paragraph was AI." There may have also been specific aspects of poorly-written essays that served as indicators of student writing, such as vague sentences (e.g., "Literature matters because literature can create a stronger message that comes across with the reader"). It is also possible that participants viewed idiosyncratic language as an indicator that the essay was written

by a student (e.g., “Today, anyone can go into a library and read about history or art and grow their brains”).

There were also cues in the writing generated by ChatGPT that seemed to influence participants’ responses. In open-ended comments, several teachers spontaneously noted that they treated the word “overall” as an indicator that the essay was generated by ChatGPT. One teacher who got a perfect score noted, “The chatbot tends to use a lot of transitional words (firstly, additionally, etc.)” However, it is important to keep in mind that not all of the heuristics teachers use are necessarily helpful [43].

We also explored teachers’ and students’ attitudes about ChatGPT. We found that teachers were less likely than students to focus on the positive implications for education, and they reported more pessimism than optimism. In contrast, students reported roughly the same level of pessimism and optimism. Students also rated various uses of ChatGPT in academic contexts as less bad. For example, although both groups thought that it was bad to submit an essay that was generated by ChatGPT for an assignment, teachers judged it to be worse. The two groups also held different views about using ChatGPT to improve an essay’s vocabulary before submitting it: on average, students viewed this as good, and teachers viewed it as bad. Future work will be needed to understand the reason for this difference. Such work could draw on the extended Technology Acceptance Model [21, 23, 44], which examines factors such as the perceived usefulness of specific technologies and how much users report trusting them. It will also be important to explore the effects of people’s prior experiences with related technologies [45].

This research has a number of limitations. We tested only a narrow range of essays in one subject area, and more work is needed to determine the generalizability of the results. It is also unclear how certain methodological decisions affected the results, such as our decision to use a word processor to correct obvious spelling and grammar errors in the student essays. In addition, it will be important to examine how providing ChatGPT with additional information in the prompt affects performance (e.g., asking it to write like a high school student or avoid certain transition words).

It is important not to overinterpret our findings. For example, the fact that experience with using ChatGPT did not appear to make a difference in accuracy should not be interpreted to mean that experience with ChatGPT cannot be used to improve one’s ability to identify writing samples that have been produced by ChatGPT. It may be that specialized training can be developed to help people pick up on cues associated with its writing style. Special software is now being developed for this purpose, but this will not be a panacea given that there will always be false positives and false negatives [46, 47]. Moreover, focusing on superficial indicators of AI-generated writing instead of learning outcomes could lead to a counterproductive “arms race” in which students try to modify their writing styles to seem less AI-like in response to teachers policing them.

Our research shows that distinguishing student writing from writing produced by chatbots such as ChatGPT is a significant challenge already, and it is likely to get even more

difficult as AI continues to improve. Our findings also suggest that content area expertise and familiarity with ChatGPT are not sufficient to overcome this problem. We suspect that total bans on ChatGPT will be ineffective at best and that a more fruitful approach will be to find ways to understand and make use of the benefits that ChatGPT offers while minimizing its risks.

Teachers who are already overburdened should not be asked to solve this problem on their own. Instead, researchers and educators will need to work together to develop best practices and effective strategies for disseminating them. Conversations that relate to this challenge are already underway, with much of the focus on using ChatGPT as a tool to achieve other goals, such as supporting language learners and providing personalized feedback on writing assignments [48]. Given that learning to work with AI systems will be a valuable skill for the foreseeable future, it will also be important to focus on teaching students how to use chatbots effectively while also thinking critically about the output they generates. As noted by New York Times tech columnist Kevin Roose, students “need to know their way around these tools — their strengths and weaknesses, their hallmarks and blind spots — in order to work alongside them” [42].

Understanding how people think about the role of ChatGPT in educational contexts is important not only due to its implications for student learning but for broader moral questions as well, including fairness. As one student commented, “I’m concerned because people who don’t use AI, like me, will see their education and grades suffer if it becomes normal.” These kinds of risks, along with the potential benefits, require creative problem-solving to effectively address. Empirical evidence and vigorous debate will play a critical role in navigating this new educational landscape.

6. Conclusion

The present research found that high school teachers were slightly better than high school students at distinguishing between essays generated by ChatGPT versus high school students. However, even the teachers had considerable trouble with this task, especially when the students’ essays were well written. Teachers who were more confident in their ability to recognize ChatGPT-generated writing did no better than their counterparts who were less confident, and experience with ChatGPT and content-relevant expertise was unrelated to this ability as well. Our attitude assessments showed that high school students had a generally negative view of submitting essays written by ChatGPT, but they viewed doing this, as well as other potential violations of academic integrity, less negatively than the teachers did. Students were more likely than teachers to report being optimistic that ChatGPT will have benefits for education. Taken together, our findings suggest that ChatGPT does pose a threat to academic integrity in high school, one that calls for the development of new policies and social norms. Given that AI will only become more sophisticated in the future, our findings raise an array of important questions about how educational systems can adjust to this powerful and rapidly advancing technology.

Data Availability

The data are available at <https://osf.io/8qv6z/>.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was supported by a grant from the National Science Foundation to T. Waltzer (#2104610). We thank Daniel Chen for assistance with study design, stimulus preparation, and pilot testing; Michael Hogan for assistance with choosing the essay prompts and for assigning the essays to students in four English classes; and Paulina Gonzales, Carmen Lonergan, and Melissa Williams for administering the survey to students in their classes. We also thank Mary Sue Compton and Mickey Freeman for rating the essays during the stimulus preparation phase and Brian Compton for comments on the manuscript. Address correspondence to G. Heyman at gheyman@ucsd.edu.

References

- [1] I. Carvalho, S. Lopes, A. Madeira, T. Palrão, and A. S. Mendes, "Robot Coworkers: The Vision of Future Hoteliers," *Human Behavior and Emerging Technologies*, vol. 2022, article 8567289, pp. 1–13, 2022.
- [2] M. Aljanabi and ChatGPT, "ChatGPT: future directions and open possibilities," *Mesopotamian Journal of CyberSecurity*, vol. 2023, pp. 16–17, 2023.
- [3] D. A. Susskind, *World without Work: Technology, Automation and how we Should Respond*, Allen Lane, London, 2020.
- [4] L. Weidinger, J. Mellor, M. Rauh et al., "Ethical and social risks of harm from language models," 2021, <https://arxiv.org/abs/2112.04359>.
- [5] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," *Nature Medicine*, vol. 25, no. 1, pp. 30–36, 2019.
- [6] H. C. Lin, Y. F. Tu, G. J. Hwang, and H. Huang, "From precision education to precision medicine," *Educational Technology & Society*, vol. 24, no. 1, pp. 123–137, 2021.
- [7] S. S. Biswas, "Role of chat GPT in public health," *Annals of Biomedical Engineering*, vol. 51, no. 5, pp. 868–869, 2023.
- [8] G. Viglia, "How to enhance critical thinking of students, educators, and researchers in the ChatGPT era," in *So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges, and implications of generative conversational AI for research, practice, and policy* International Journal of Information Management in press.
- [9] R. Luckin, W. Holmes, M. Griffiths, and L. B. Forcier, *Intelligence Unleashed: An Argument for AI in Education*, Pearson, London, 2016.
- [10] S. Akgun and C. Greenhow, "Artificial intelligence in education: addressing ethical challenges in K-12 settings," *AI and Ethics*, vol. 2, no. 3, pp. 431–440, 2022.
- [11] N. Selwyn, "The future of AI and education: some cautionary notes," *European Journal of Education*, vol. 57, no. 4, pp. 620–631, 2022.
- [12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," 2018, <https://arxiv.org/abs/1810.04805>.
- [13] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [14] T. Susnjak, "ChatGPT: the end of online exam integrity?," 2022, <https://arxiv.org/abs/2212.09292>.
- [15] D. Cotton, P. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT," 2023, <https://edrxiv.org/mrz8h/>.
- [16] A. Mitchell, "Professor catches student cheating with ChatGPT: 'I feel abject terror'," *New York Post*, 2022, <https://nypost.com/2022/12/26/students-using-chatgpt-to-cheat-professor-warns>.
- [17] C. Stokel-Walker, "AI bot ChatGPT writes smart essays — should professors worry?," *Nature*, 2022.
- [18] T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing, "Exploring AI ethics of ChatGPT: A diagnostic analysis," 2023, <https://arxiv.org/abs/2301.12867>.
- [19] D. Rosenzweig-Ziff, "New York City blocks use of the ChatGPT bot in its schools," *The Washington Post*, 2023, <https://www.washingtonpost.com/education/2023/01/05/nyc-schools-ban-chatgpt/>.
- [20] I. Celik, "Towards intelligent-TPACK: an empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education," *Computers in Human Behavior*, vol. 138, article 107468, 2023.
- [21] R. Chocarro, M. Cortinas, and G. Marcos-Matás, "Teachers' attitudes towards chatbots in education: a technology acceptance model approach considering the effect of social language, bot proactiveness, and users' characteristics," *Educational Studies*, vol. 49, no. 2, pp. 295–313, 2023.
- [22] X. Zhai, X. Chu, C. S. Chai et al., "A review of artificial intelligence (AI) in education from 2010 to 2020," *Complexity*, vol. 2021, Article ID 8812542, 18 pages, 2021.
- [23] S. Kelly, S.-A. Kaye, and O. Oviedo-Trespalacios, "A multi-industry analysis of the future use of AI chatbots," *Human Behavior and Emerging Technologies*, vol. 2022, article 2552099, pp. 1–14, 2022.
- [24] J. A. Lee, A. Bardi, P. Gerrans et al., "Are value-behavior relations stronger than previously thought? It depends on value importance," *European Journal of Personality*, vol. 36, no. 2, pp. 133–148, 2022.
- [25] E. Turiel, "Morals, motives and actions," *British Journal of Developmental Psychology Monograph Series*, vol. 2, pp. 29–40, 2003.
- [26] T. Waltzer and A. Dahl, "Why do students cheat? Perceptions, evaluations, and motivations," *Ethics & Behavior*, vol. 33, no. 2, pp. 130–150, 2023.
- [27] J. J. Arnett, "Conceptions of the transition to adulthood: perspectives from adolescence through midlife," *Journal of Adult Development*, vol. 8, no. 2, pp. 133–143, 2001.
- [28] S. Jiang, S. D. Simpkins, and J. S. Eccles, "Individuals' math and science motivation and their subsequent STEM choices and achievement in high school and college: a longitudinal study of gender and college generation status differences," *Developmental Psychology*, vol. 56, no. 11, pp. 2137–2151, 2020.
- [29] V. E. Gunser, S. Gottschling, B. Brucker, S. Richter, and P. Gerjets, "Can users distinguish narrative texts written by an artificial intelligence writing tool from purely human text?,"

- in *HCI International 2021 - Posters*, C. Stephanidis, M. Antona, and S. Ntoa, Eds., vol. 1419, pp. 520–527, Springer, 2021.
- [30] N. Köbis and L. D. Mossink, “Artificial intelligence versus Maya Angelou: experimental evidence that people cannot differentiate AI-generated from human-written poetry,” *Computers in Human Behavior*, vol. 114, article 106553, 2021.
- [31] W. Yeadon, O. O. Inyang, A. Mizouri, A. Peach, and C. Testrow, “The death of the short-form physics essay in the coming AI revolution,” *Physics Education*, vol. 58, no. 3, article 035027, 2023.
- [32] C. Terwiesch, *Would Chat GPT3 Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course*, Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania, 2023, <https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP-1.24.pdf>.
- [33] OpenAI, “GPT-4 technical report,” 2023, <https://arxiv.org/abs/2303.08774>.
- [34] J. C. F. de Winter, *Can ChatGPT Pass High School Exams on English Language Comprehension?*, Delft University of Technology, 2023.
- [35] S. Erickson and E. Heit, “Metacognition and confidence: comparing math to other academic subjects,” *Frontiers in Psychology*, vol. 6, p. 742, 2015.
- [36] I. Fischer and D. V. Budescu, “When do those who know more also know more about how much they know? The development of confidence and performance in categorical decision tasks,” *Organizational Behavior and Human Decision Processes*, vol. 98, no. 1, pp. 39–53, 2005.
- [37] J. T. Wixted and G. L. Wells, “The relationship between eyewitness confidence and identification accuracy: a new synthesis,” *Psychological Science in the Public Interest*, vol. 18, no. 1, pp. 10–65, 2017.
- [38] C. Clerwall, “Enter the robot Journalist,” *Journalism Practice*, vol. 8, no. 5, pp. 519–531, 2014.
- [39] T. Waltzer and A. Dahl, “Students’ perceptions and evaluations of plagiarism: effects of text and context,” *Journal of Moral Education*, vol. 50, no. 4, pp. 436–451, 2021.
- [40] T. Waltzer, F. C. DeBernardi, and A. Dahl, “Student and teacher views on cheating in high school: perceptions, evaluations, and decisions,” *Journal of Research on Adolescence*, vol. 33, no. 1, pp. 108–126, 2023.
- [41] J. Hox, M. Moerbeek, and R. van de Schoot, *Multilevel Analysis: Techniques and Applications*, Routledge, New York, NY, 3rd edition, 2010.
- [42] K. Roose, *Don’t ban ChatGPT in schools. Teach with it*, The New York Times, 2023, <https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html>.
- [43] M. Jakesch, J. T. Hancock, and M. Naaman, “Human heuristics for AI-generated language are flawed,” *Proceedings of the National Academy of Sciences*, vol. 120, no. 11, article e2208839120, 2023.
- [44] F. D. Davis, “Perceived usefulness, perceived ease of use, and user acceptance of information technology,” *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.
- [45] C. Shao and K. H. Kwon, “Hello Alexa! Exploring effects of motivational factors and social presence on satisfaction with artificial intelligence-enabled gadgets,” *Human Behavior and Emerging Technologies*, vol. 3, no. 5, pp. 978–988, 2021.
- [46] Y. Barlette, “ChatGPT in the context of education: How to deal with its disruptive effects,” in *“So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy*, International Journal of Information Management in press.
- [47] S. Svrluga, “Princeton student builds app to detect essays written by a popular AI bot,” The Washington Post, 2023, <https://www.washingtonpost.com/education/2023/01/12/gptzero-chatgpt-detector-ai/>.
- [48] R. Raman, G. H. Davies, and A. Mardani, “ChatGPT and education: Is a symbiotic relationship possible?,” in *“So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy*, International Journal of Information Management in press.