The Unexpected Effects of Google Smart Compose on Open-Ended Writing Tasks

Robert E. Cummings[1][0000−0002−0390−3710], Thai Le[1][0000−0001−9632−6870], Sijan Shrestha[1], and Carrie Veronica Smith[1][0000−0002−9123−0244]

University of Mississippi, University MS 38677-1848, USA

Abstract. In this work we seek to understand the effects of Google Smart Compose's auto-complete suggestions on writing products and processes in open-ended writing tasks. We recruited 119 human subjects to a carefully designed user study to write timed responses to a common prompt via word processors while recording interactions with Google Smart Compose behind the scene. Our experiment demonstrates the impacts of text suggestions on writing length, structure, cohesion, and complexity, utilizing Coh-metrix values. We also introduce new metrics for measuring and understanding detailed human subject interactions with text suggestions. We find that writers utilizing Google Smart Compose write *the same amount of text* as those without Smart Compose, and that the structure of written work produced with Google Smart Compose enabled is *not significantly distinguishable* from writing produced with Google Smart Compose disabled. In addition, there was no strong evidence that writing process was significantly impacted by Google Smart Compose. Finally, this work identifies factors that can be used in future studies measuring language model interactions with human writers.

Keywords: Auto Completion · AI-Assisted Writing · Google Smart Compose · Predictive Text Suggestions

1 Introduction

The release of GPT-3 in 2020 triggered a dramatic increase in popular awareness of the ability of auto-regressive large language models (LLMs) to produce text to assist human writing tasks. The class of tools produced with GPT-3 and its successors such as ChatGPT (GPT-3.5) and GPT-4, also known as AI-powered writing generators, provide strings of text in response to user inputs. These outputs are longer and more sophisticated than earlier AI-powered writing tools classified as writing assistants, which include next-word-prediction and sentence completion technologies such as Google Smart Compose (GSC) [19]. As writing generators have proliferated, so too has the range of their writing goals, including summarizing texts [5], writing social media posts [9], finding research to support assertions [15], and even creative writing [22, 7].

As AI-powered writing generators are most often deployed with common web searches and office productivity softwares with little or no user training, the

education community is increasingly aware of the need for intentional and structured engagements with these tools. Educators understand that risks for using AI tools include "plagiarism, harmful and biased content, equity and access, the trustworthiness of the AI-generated content, and over-reliance on the tool for assessment purposes" [20]. The effective use of AI-powered writing generators also depends greatly upon their reception by human writers. Understanding the context, purpose, genre, and technological platform utilized by human writers is vital to shaping the value of AI outputs and how human writers seek to incorporate AI writing outputs within writing projects. Consequently, researchers are also paying greater attention to the context and reception of AI writing outputs as they develop their tools [5]. Further, researchers have now developed more comprehensive frameworks for studying and engaging human reception for AI writing outputs [10].

A long term claim for the value of similar artificial intelligence (AI) technologies has been to increase human productivity [21, 3]. The developers of the current AI-powered writing generators, including ChatGPT, offer similar claims [18]. But the longer established AI-powered writing applications have not yet been thoroughly examined to determine if claims of increasing writing productivity are true. These technologies include applications that mainly focus on next-word-prediction or sentence completion rather than generating the whole sentences for the users.

Although recently there is great interest in the emergence of LLMs applied to digital writing processes, precedent softwares such as Grammarly and GSC remain understudied in relation to their broad adoption. This lack of understanding has a direct impact on pedagogical strategies within education systems for the understanding and use of AI-powered writing generators. Currently, many higher education faculty are experimenting with AI-powered writing generators in their classrooms with little guidance from the research community [17]. In order to train students to effectively integrate these tools in their writing processes for both greater productivity and informed ethical awareness, faculty will *need a greater understanding of the potential impact of AI-powered writing generators on their students' writing.*

We believe that a more comprehensive understanding of human-computer interactions around the utilization of these technologies is a necessary building block to understanding the more recent developments in applied LLMs to writing assistant technologies. Therefore, in this work, we propose to investigate the effects of the popular writing assistant technology GSC on open-ended writing tasks. This work, being conducted in an academic setting with human subjects writing in an academic genre, will informative for students, faculty, and administrators seeking to understand and utilize AI-powered writing generators in the classroom.

2 Related Work

Google Smart Compose. GSC was first launched in August of 2019 as an assistive tool for G-mail [2]. Later in 2019, Smart Compose was also offered within Google Docs [19]. As a writing assistance technology, GSC is an extension of earlier predictive text suggestion technologies developed for mobile platforms that attempt to assist writers in composing texts by suggesting the completion of words, phrases, and sentences. These technologies differ significantly based on their platform (i.e., mobile, web-based, word processor embedded) and their related genre (e.g., texting, e-mail, open-ended writing). GSC is now a widely adopted AI-powered writing assistant, and part of a class of technologies generally known as predictive or suggestive text writing assistants that offer writers the opportunity to complete statements with text strings supplied by computer applications. Smart Compose represents only a recent development in the steady rise of automated digital writing assistants, starting with Word processing spell checkers, grammar correction technologies, and now LLMs [4].

Human Study on Writing Assistants. Prior studies have examined human computer interactions in email composition [2] and in mobile devices [16]. In particular, Arnold et al., demonstrated that writing assistant technologies have the potential to strongly influence the output of human writers in terms of content, diction, syntax, and linguistic complexity [1]. However, prior studies have been conducted only in narrow writing tasks, such as determining one word, phrase, or sentence to describe an image. Thus, we seek to better understand the effects of writing assistant technologies when applied to more *open-ended writing tasks*, conducted in word processors on laptops, where writers have *nearly unlimited writing freedom to reach a goal of expression that is more nuanced and complex, and typical of higher stakes writing tasks.*

3 Research Questions

Our main research focus is measuring the impact of GSC both on **writing product and on writing process** with writers who are given an open-ended writing task, utilizing word processors on laptop computers. Our single study variable was whether or not GSC was enabled. Since we are able to compare writing samples and writing experiences of writers responding to the same prompt, either with or without GSC enabled, we formulate the following research questions (RQs) around writing product and processes:

- **RQ1 (Writing Product)**: *Are writing products composed with Google Smart Compose different from those composed without Google Smart Compose in terms of length and textual complexity?*
- **RQ2 (Writing Process)**: *How does Google Smart Compose affect human writing behaviors such as accepting, editing, and rejecting suggestions?*
- **RQ3 (Time Spent with Google Smart Compose)**: *Does time spent on writing suggestions correlate with different behavior patterns such as accepting, editing, or rejecting suggestions?*

In **RQ1**, we wished to understand the depth of the impact of GSC suggestions not only on the length of the writing samples, but also on upon the complexity or depth of written expression. We wondered if suggestions offered by GSC might be similar across writing sessions where it was enabled, and this push written expression to be more similar across samples from those sessions. Beyond the writing products themselves, in **RQ2**, we were interested in better understanding how GSC text suggestions affected human writing processes or behaviors. To gain more insight into the writing process of writers who utilize GSC, we devised three new measures to record how writers responded to GSC suggestions, tracking whether writers (1) accepted the suggestion verbatim (full acceptance); (2) edited a suggestion before accepting it (partial acceptance), or (3) rejected the suggestion outright (rejection) and also how much time it took for them to make those decisions. In **RQ3**, we are also interested to know the time spent reading, evaluating, revising, and/or ultimately rejecting Google Smart Compose suggestions and their comparisons.

4 User Study

4.1 Recruitment, Consent and Compensation

We recruited 119 participants at a US public research university to write for 25 minutes in response to a prompt that solicited their opinions about the roles of luck and/or hard work in accounting for success in life. No university status was sought or required (i.e., no student status was required), no literacy ability was measured, and all participants were 18 years of age or older. We randomly enabled/disabled GSC for each participant via uniform sampling. This resulted in roughly half (55) of the participants writing with GSC enabled, and roughly half (59) writing without GSC enabled. All participants wrote on laptop computers in a room with up to four other participants, with access to the room controlled by a research assistant. The 25 minute time limit for sample collection was monitored by our custom made software, and the research assistant guided participants to and from laptop work stations after participants had completed the consent process.

During the consent process, study participants were initially deceived as to our interest in measuring the impact of GSC on writing. At the conclusion of the writing experiment participants were informed of the purpose of the study and asked to re-consent to the collection of their data. Participants were compensated 25.00 USD (about $60.00/hour compared to the current federal minimum wage of $7.25) for their time regardless of whether they consented to allow the collection, analysis, and reporting of their data. Our study was reviewed and received exempt status from our institutional IRB office.

4.2 Participants and Demographics

We collected demographic information on our study participants. This included our participants' genders (59% Male, 39% Female, 2% Non-Binary), the age of

our participants (58% 18-22, 29% 23-29, 10% 30-39 3% 30-49) and our participants' racial and ethnic identities (49% Caucasian, 12% African-American, 3% Latino or Hispanic, 29% Asian, 3% Other or unknown and 1% No Response). We also queried our participants about their highest level of education completed (3% PhD, 17% Master's degree, 24% Bachelor's degree, 16%, 15% and 15% are in third, second, first year of college, 10% high school).

After recording demographic information for our study participants, we also asked two questions about participants' experience with auto-complete technologies generally (multiple responses allowed). 75%, 67% and 61% reported to have some experience with "word processor", "texting or mobile applications", "web-based applications", respectively. We also asked their frequency of the technology usage: 33% responded "very frequently", 25% "frequently", 25% "sometimes", and 13% responded "rarely". Only 1% responded "never" and 3% did not give any responses. We also asked a final question about users' opinions of auto-complete technologies, where 37% and 36% responded to "very favorable", "favorable", 19% responded to "neutral", 3% responded to "negative", none responded to "very negative" and 5% did not give any responses. For the purpose of the survey, we defined auto-complete technologies as "any technology that suggests how you should complete sentences or words on any digital writing platform."

4.3 Apparatus

We developed a custom user-interface to enable our research. GSC is only available via Google Docs web application, a well-known online collaborative text processor. Since this auto-suggestion technology is provided as an integrated function of Google Docs *rather than a public API access*, it is non-trivial to systematically capture all the needed interactions between the study participants and the software for our study. To resolve this, we developed a custom web browser on top of the the open-sourced Google Chrome driver[1] in Python programming language. Although this browser looks exactly the same as the official Google Chrome browser, by using it to access Google Docs, we are able to extract all the necessary information to track suggestions offered to participants. This information includes each input keystroke from the users, the writing after each additional keystroke, suggested phrases from GSC, the prompts and keystrokes that triggered them, and the final writing products. Importantly, this information also includes respective timestamps, including when a phrase is suggested by GSC and when the user takes the next action. Using this raw data, we are then able to derive useful statistics to answer the proposed research questions.

4.4 Study Design

Our study design is a *post-test only* experimental design. All participants were instructed to write for 25 minutes on the topic of, *"What do you think is more*

[1] https://chromedriver.chromium.org/

important for success in life: luck or hard work?" Participants wrote their open-ended responses on provided laptop computers, as more fully described in section 4.3. Participants were randomly assigned to two groups – a control group that wrote as normal (GSC disabled) and an experimental group, whose computer was enabled with GSC. The independent variable was the presence of GSC (yes/no). The dependent variables were multiple indices and measures capturing both writing production and writing process, which we will describe below.

4.5 Measures

To structure our results reporting and subsequent analysis, we split our concerns into the impact of GSC on *writing product*, or the text, and *writing process*, or the actions of writers engaging their writing technologies to create the finished documents.

Writing Product. To analyze the writing products, we deployed *Coh-Metrix 3.0*, an automated tool for text and discourse analysis [12]. Coh-metrix is a good fit for our research because it is a well-established tool that can provide 108 quantitative and linguistic measurements for the writing samples. These measurements include a range of dimensions: basic textual descriptions (e.g., number of words in a writing sample, number of sentences, number of paragraphs), textual cohesion, lexical complexity, syntactic complexity, reading difficulty, and additional measurements. Given the fact that Coh-Metrix has been used continuously in some form since 2004 [8], broadly adopted (thousands of peer-reviewed studies utilize some versions of the tool), validated [13], widely used by the NLP community [11], and can help us to explore multiple dimensions of writing products simultaneously, we chose it as the primary tool for our textual analyses.

Writing Process. To analyze writing processes, we devised three new measures to record how writers responded to the suggestions of GSC, tracking whether writers (1) accepted the suggestion verbatim; (2) edited a suggestion before accepting it, or (3) rejected the suggestion outright. If a writer accepted a GSC suggestion as it was offered, we coded that action as "full acceptance," or FA. If a writer edited the GSC suggestion before accepting it, we coded that action as "partial acceptance," or "PA". And if a writer completely rejected a GSC suggestion, we coded that action as "Rejection." We tracked all GSC suggestions equally, regardless of whether would suggest altering a word, phrase, or sentence. In addition, we tracked whether a writer "backtracked," or later returned to edit a suggestion that had been addressed earlier.

After defining these three possibilities, we tracked how each writer deployed the three options, in order to build profiles of each writer's choices. For each category (Full Acceptance, Partial Acceptance, Rejection), we recorded all statistics as described in Table 1C&D.

4.6 Analysis Methods

To best describe the scope of the writing samples provided by our participants, we engaged the several descriptive indices of writing samples from Coh-Metrix

Table 1. Statistical measures for measuring writing product and writing process

Writing Product
(A) Scope of the writing samples: # of paragraphs (DESPC), # of sentences (DESSC), # of words (DESWC), mean # of words found in each sentence (sentence length) (DESSL)
(B) Reading, structure and lexical complexity: Flesch Reading Ease (RDFRE), Syntactic simplicity (PCSYNz), Syntactic complexity – words before main verb (SYNLE), Lexical diversity – content words (LDTTRc)
Writing Process
(C) Individual writer: # of times engaged in an action, % of the three possible actions taken, ET on an action, % of ET on an action out of all possible responses
(D) Individual writer v.s. all writers: Rank of (1) against AW, % of three possible actions by AW, % of total ET on three possible actions by AW, % of ET on an action compared to total ET on accepts by AW, Rank of total ET spent on accepts as compared to AW

ET: Elapsed Time Spent; AW: all writers with GSC enabled

as shown in Table 1A, including measures such as number of words, paragraphs and sentence lengths. [2] We selected these measures as they provided a general sense of the scope of the writing samples, and enabled comparisons between when GSC disabled and enabled writing samples. Particularly, we were interested in understanding how the suggestions of GSC could affect these values because we reasoned that writers who frequently accept suggestions verbatim (FA) might display higher word counts than those who did not. Similarly, we reasoned that writers who heavily edited suggestions might write less than those who did not, in part because more of their allowed time would be spent on editing as opposed to generating text. Similarly, we reasoned that if GSC suggestions perpetually presented writers with a means to end a sentence quickly, then sentence length could be affected when comparing writing samples produced with and without suggestions.

In order to facilitate an understanding of the different levels of reading complexity, textual coherence, syntactic complexity, and lexical complexity between writing samples, we applied additional Coh-Metrix measures as described in Table 1B, including reading ease, syntactic simplicity and complexity and lexical diversity. Additionally, we looked for to the demographic information collected about our participants (Sec. 4.2). In particular, we looked at how frequently our participants indicated that they used auto-complete technologies (1=rarely used, 5=very frequently used).

[2] Full definitions of these measures can be found in [12]

Table 2. Quantitative and qualitative statistics of writing production

	Smart Compose Disabled (n=59)				Smart Compose Enabled (n=55)			
	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD
Quantitative Indices								
Number of paragraphs	1	14	4.39	2.91	1	10	3.84	2.47
Number of sentences	11	60	29.66	11.14	14	77	28.96	10.88
Number of words	303	1,096	595.56	190.42	241	1,245	585.85	205.05
Number of words/sentence	13.21	32.27	21.00	4.50	12.61	42.00	21.08	6.01
Qualitative indices								
Syntactic simplicity	-1.57	1.06	-0.30	0.60	-2.02	1.12	-0.32	0.61
Syntactic complexity	2.09	9.00	4.39	1.31	2.25	7.56	4.16	1.38
Lexical diversity	.40	.79	.62	.08	.50	.79	.63	.07
Reading ease	39.94	82.14	67.02	8.09	37.72	80.68	65.62	9.09
Referential cohesion	-1.06	2.19	0.41	.71	-1.10	2.14	0.36	0.75

5 Results

We engaged 119 participants, but due to unforeseeable technical difficulties, only 114 writing samples were usable. Of those 114 samples, 59 were written with GSC disabled, and 55 were written with GSC enabled.

5.1 Impact of Google Smart Compose on Writing Product (RQ1)

To ensure the effectiveness of random assignment, we conducted an independent samples t-test to determine whether the experimental and control groups differed on typical usage of auto-complete technology. The experimental (M=3.89, SD=1.17) and control (M=3.70, SD=1.05) groups did not significantly differ, with t(108)=-0.87, ns. The means and standard deviations for the quantitative and qualitative indices of writing production are presented in Table 2.

A series of regression analyses were conducted to examine whether the usage of GSC technology resulted in quantitatively or qualitatively different writing production. In addition to the primary variable of interest, experimental condition (*enabled* or *disabled*), we also included typical usage of auto-complete technology in order to control for potential pre-existing differences between users.

Experimental condition was *not* a significant predictor of either the quantitative (see Table 3 for regression coefficients) or the qualitative indices (see Table 4 for regression coefficients) of the writing production variables, nor was self-reported auto-complete usage.

5.2 Impact of Google Smart Compose on Writing Process (RQ2)

In addition to evaluating the effect of GSC on our open-ended writing samples, our third research question invites us to consider how writers interact with GSC to produce text.

Table 3. Regression analyses predicting quantitative indices of writing production by experimental condition

	# of Paragraphs			# of Sentences			# of Words			# of Words/Sentence		
	B	β	t	B	β	t	B	β	t	B	β	t
Intercept	5.74			30.78			635.12			22.96		
Condition	-.36	-.15	-1.53	-.27	-.03	-0.28	-9.55	-.05	-0.56	.07	.007	0.07
Usage	-.47	-.09	-0.90	-.63	-.03	-0.29	-10.48	-.03	-0.28	-.50	-.11	-1.10

Table 4. Regression analyses predicting qualitative indices of writing production by experimental condition

	Syntactic Simplicity			Syntactic Complexity			Lexical Diversity			Reading Ease			Referential Cohesion		
	B	β	t	B	β	t	B	β	t	B	β	t	B	β	t
Intercept	-0.52			4.62			.62			66.41			066		
Condition	-.03	-.03	-0.28	-.28	-.10	-1.06	.003	.02	0.22	-1.22	-.07	-0.73	-.03	-.02	-0.23
Usage	.06	.11	0.11	-.05	-.04	-0.45	.002	.03	0.27	.15	.02	0.19	-.07	-.10	-1.05

The 55 participants in the GSC enabled group received an average of 68.45 auto-complete recommendations, with a mode of 73. The number of auto-complete recommendations ranged from 12 to 146, with a standard deviation of 34.13. Because writing more allows for more recommendations, correlations between the number of recommendations and quantitative indices of writing production are uninformative although we did calculate these analyses and all correlations were significant and positive, $r's \geq 0.39$, with the exception of mean number of words per sentence, which approached conventional levels of significance ($r(53)=0.25, p=0.07$). However, receiving more recommendations may result in better writing quality. A series of regression analyses was conducted to examine this question, with writing quality metrics regressed on number of recommendations received and typical usage of auto-complete (entered as a control variable). Results suggest that receiving more recommendations was not significantly associated with any of the qualitative indices of writing production.

To determine if participants were equally likely to handle recommendations in the same way, a one-way repeated measures ANOVA was performed. Mauchly's test indicated that our data did not violate the assumption of sphericity, $W=0.94$, $\chi^2(2)=3.11$, ns. Significant differences were found in the participants' tendencies to fully accept, partially accept, and reject recommendations, $F(2, 108)=95.485$, $p<.001$. Post-hoc tests show significant differences in between the three decisions. Specifically, participants accepted more recommendations ($M=32.36$, $SD=16.59$) than they rejected ($M=22.36$, $SD=22.36$) and partially rejected ($M=13.73, SD=8.11$, $ps<.001$). Given that recommendations are meant to assist the writer, a series of regression analyses examined whether proportion of full acceptances (or receiving recommendations the participant found helpful,

relative to the number of total recommendations received) predicted differences in writing production, quantitatively or qualitatively. Again, typical usage of auto-complete technology was added to the model as a control variable. *Receiving helpful recommendations was not significantly associated with any metric of writing production, either qualitative or quantitatively.*

5.3 Examining Time Spent with Google Smart Compose Recommendations (RQ3)

The 55 participants in the GSC enabled group spent an average of 12,420 milliseconds interacting with recommendations. The amount of time spent with recommendations ranged from 2,628 to 26,635 milliseconds, with a standard deviation of 5,621 milliseconds. To determine whether participants spent equal amounts of time with each type of recommendation (acceptance, partial acceptance, and rejection), a one-way repeated measures ANOVA was performed. Mauchly's test indicated that our data did violate the assumption of sphericity, W=0.71, $\chi^2(2)$=18.35, p<.001. Since the epsilon value was less than 0.75, a Greenhouse-Geisser correction was used. Significant differences were found in the amount of time participants spent with each type of recommendation, $F(1.55, 83.55)$=138.07, p<.001. Post-hoc tests indicated significant differences between the three; participants spent significantly more time with acceptances (M=7, 524.26, SD=45.20) than they did with partial acceptances, (M=3, 160.99, SD=300.98) and rejections, (M=1, 734.26, SD=170.59) and more time with partial acceptances than they did with rejections, ps<.001.

6 Discussion

6.1 AI in Education Implications

This research has direct implications for the use of AI in education for several reasons. First, the writing genre of our experiment is an open-ended argument. Writers were asked to offer a clear statement in response to our prompt, and were also requested to provide evidence in support of that statement. Argument is a common mode of writing throughout secondary and higher education systems and is used in both high-stakes and low-stakes assignments [14]. Therefore any observations of how writers utilize AI-powered writing generators within the genre of argument are more likely to be correlate with the everyday writing tasks of higher education classrooms. Secondly, our research was conducted on a university campus with human subjects. While we cannot state that every participant in our study was a college student, we do know that all of them attained education beyond high school and 58 percent were within the age range of traditional undergraduates. Others could have also been graduate students.

Therefore, given the ages, educational attainment, and setting of our experiment, the results from our experiment are more likely to duplicate the experience of higher education students utilizing AI-powered writing generators for

academic assignments in higher education classrooms. While we are aware of recent research that has compared the reception of sentence level text suggestions to message level text suggestions with AI-powered writing generators, we note that this research was conducted with online participants writing in fictional roles. [6] We believe that this work with human subjects, writing about their genuine beliefs, and supporting those opinions with first-hand evidence, is more likely to duplicate the higher education writing experience.

6.2 Unexpected Results regarding Google Smart Compose's Effects

Results regarding the effects of GSC were unexpected. We created Coh-Metrix reports for all 114 usable writing records. In reviewing the results to answer **RQ1**, we did not find a significant difference in writing sample length when comparing the average number of words written with GSC enabled versus disabled (see Table 2). Indeed, the length of writing samples produced without GSC were nearly identical in word length to those produced with GSC. Writers with GSC enabled wrote an average of only 2% less words than writers without GSC enabled, and virtually the same number of sentences.

Moreover, we hypothesized that writing samples produced with GSC enabled would be characterized by lower textual complexity than their counterparts written without GSC. We reasoned that when writers are utilizing uniform versions of Smart Compose that have not been customized to the writer, text suggestions would also be more uniform, and if accepted, would lead to more similarity in the written product. Additionally, since a GSC text suggestion will almost always suggest a way to complete a sentence, the cumulative effect could lead to writing samples that had shorter sentence length. Here again, the results of our study clearly demonstrate that GSC had no meaningful impact on textual complexity, when examined in terms of lexical complexity, syntactic complexity, or reading ease.

Our subsequent **RQ2**, were all three questions of writing process. Each looked at how a group of writers who spent more time with particular category of GSC suggestion – "full acceptance," "partial acceptance" and "rejection" – might have writing products or processes that stood out from the others. During our study we collected a substantial amount of writing process data. Although we have not exhausted every possible inquiry of those data, there are no preliminary identifiable differences in the writing products between writers who heavily engaged in one of those strategies.

The most striking finding of our work is the lack of impact that GSC had upon both writing product (the text itself) and the writing process (the approach of writers in creating text). The written work of participants using GSC was almost indistinguishable from work produced without GSC. We did note that timing measurements indicate that writers who had GSC enabled took longer to accept suggestions as offered than to either edit or reject them. This finding has implications for how we should characterize writers' responses to text suggestions, and implies that writers are investing the most time and thought when evaluating a machine-generated suggestion for full acceptance, as compared to editing

or rejecting suggestions. This raises questions about the investment of writers in their own words when compared to text that is offered via AI, and has potential implications for further studies, including those that would record the whether suggestions increase speed without decreasing quality.

Similarly, when we examined demographics, we did not find a significant correlation between participant familiarity with auto-complete technologies and writing product, nor did we find any correlation with participants attitudes toward the technologies and their impact on writing products or processes. To guard against writing abilities being overly weighted in either experiment or control groups, we did review the highest level of education completed in both groups, deeming this as a reasonable proxy for writing abilities. While the experiment group did reveal a slightly higher concentration of graduate students when compared to the control group (29% to 18%), we do not believe them to be overly represented with a material consequence.

7 Limitations and Future Works

7.1 Limitations

We acknowledge that our study utilized instances of GSC that were "off-the-shelf." That is, when we deployed Smart Compose for our subjects, each use of the software was new – the software was not able to customize its suggestions to users based on a developed history of user preferences. By deploying GSC without user customization, we defeated a potential benefit of the software for writers (i.e., more accurate word suggestions) and a presumably a primary feature intended by designers.

Our prompt was written by the researchers for this experiment only. The operational section of the writing prompt is: *"What do you think is more important for success in life: luck or hard work? And why? Feel free to support your answer with examples from your own lived experience as well as the examples of famous men and women. This includes citing to examples from your personal experience, your professional experience, and public figures we all know."* We recognize that this prompt could produce an informal style of writing, where writers were invited to identify and discuss personal experiences as well as the experiences family members. This storytelling mode could also push writers toward greater cohesion than if they were asked to write about unfamiliar and more abstract ideas.

Lastly, we failed to record participants' native language. If English is a second language for a study participant, it is possible that L2 factors could have affected writing productivity and complexity.

7.2 Future Works

There are several possible explanations for the lack of impact GSC demonstrated on both writing process and writing product that might be engaged by future studies.

Have writers have already adjusted to using auto-complete technologies? Humans do not read and write simultaneously. Our presumption in designing this study was that the cognitive change from writing to reading disrupts the writing process, and that part of that work would be measuring and describing the nature of that disruption. Further, we expected that different writers would accommodate the writing technologies differently. Although our population was small when compared to the total number of GSC users, their writing processes were clearly not diminished by the cognitive load presented by interpreting, evaluating, integrating or rejecting, auto-complete suggestions. Clearly, our findings refute the idea that auto-complete suggestions play a disruptive impact on the writing product or process, at least when utilized for open-ended writing projects conducted on laptops utilizing word processors. If writers do not find the current level of auto-complete suggestions disruptive to the composing process, what does this possibly signal for the development of future assistants?

Does the writing context have a greater impact that previously understood? Another possible explanation for our findings of little impact for GSC on the length or quality of writing could be attributed to the impact of the writing context: the writer's purpose, the technology deployed, and the genre engaged. At least one previous study of the impact of predictive text technologies on writing samples found that text suggestions held great influence on the amount of text written and the diction used [1]. However, this study looked at captions for images written on mobile platforms, where only one word or phrase was produced by writers. In contrast, our study asked writers to formulate ideas in an open-ended writing context and via word processors. Future studies that look to better understand the impact of AI-powered writing assistants (and generators) need to fully address the context the writing purpose, platform, and genre.

8 Conclusion and Future Works

This work investigates the impacts of the popular writing assistant GSC technology on open-ended writing tasks, both on writing product and on writing process. Our user study showed that writers write the same amount of text regardless of whether or not they utilize GSC, and that the structures of writing produced with GSC is not significantly distinguishable from those produced without it. Moreover, there was no strong evidence that GSC impacted the writing process. Because both the technology producing writing suggestions, and the cognitive state of the writer reviewing suggestions, are complicated and constantly in-flux throughout a writing session with GSC enabled, we believe that our study offers an important first look into an incredibly complex interchange. We believe that our key findings for writing process and writing product with text suggestions in open-ended writing environments contribute to a growing understanding of contemporary digital writing. Our work helps teachers and educators better understand and utilize both existing and future AI-powered writing assistants and generators in the classroom.

References

1. Arnold, K.C., Chauncey, K., Gajos, K.Z.: Predictive text encourages predictable writing. In: ACM IUI (2020)
2. Chen, M.X., Lee, B.N., Bansal, G., Cao, Y., Zhang, S., Lu, J., Tsay, J., Wang, Y., Dai, A.M., Chen, Z., et al.: Gmail smart compose: Real-time assisted writing. In: KDD (2019)
3. Crawford, K.: The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press (2021)
4. Dale, R., Viethen, J.: The automated writing assistance landscape in 2021. Natural Language Engineering (2021)
5. Dang, H., Benharrak, K., Lehmann, F., Buschek, D.: Beyond text generation: Supporting writers with continuous automatic text summaries. In: UIST (2022)
6. Fu, L., Newman, B., Jakesch, M., Kreps, S.: Comparing sentence-level suggestions to message-level suggestions in ai-mediated communication. In: CHI (2023)
7. Gómez-Rodríguez, C., Williams, P.: A confederacy of models: a comprehensive evaluation of llms on creative writing. arXiv arXiv:2310.08433 (2023)
8. Graesser, A.C., McNamara, D.S., Kulikowich, J.M.: Coh-metrix: Providing multilevel analyses of text characteristics. Educational researcher (2011)
9. Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., Naaman, M.: Co-writing with opinionated language models affects users' views. arXiv arXiv:2302.00560 (2023)
10. Lee, M., Srivastava, M., Hardy, A., Thickstun, J., Durmus, E., Paranjape, A., Gerard-Ursin, I., Li, X.L., Ladhak, F., Rong, F., et al.: Evaluating human-language model interaction. arXiv arXiv:2212.09746 (2022)
11. McNamara, D.S., Graesser, A.C.: Coh-metrix: An automated tool for theoretical and applied natural language processing. In: Applied natural language processing: Identification, investigation and resolution. IGI Global (2012)
12. McNamara, D.S., Graesser, A.C., McCarthy, P.M., Cai, Z.: Automated evaluation of text and discourse with Coh-Metrix. Cambridge University Press (2014)
13. McNamara, D.S., Ozuru, Y., Graesser, A.C., Louwerse, M.: Validating coh-metrix. In: Proceedings of the 28th annual conference of the cognitive science society (2006)
14. Nesi, H., Gardner, S.: Genres across the disciplines: Student writing in higher education. Cambridge University Press (2012)
15. Ought: Elicit: The ai research assistant (2023), https://elicit.org
16. Quinn, P., Zhai, S.: A cost-benefit study of text entry suggestion interaction. In: CHI (2016)
17. Tate, T., Doroudi, S., Ritchie, D., Xu, Y., et al.: Educational research and ai-generated writing: Confronting the coming tsunami (2023)
18. Teubner, T., Flath, C.M., Weinhardt, C., van der Aalst, W., Hinz, O.: Welcome to the era of chatgpt et al. the prospects of large language models. Business & Information Systems Engineering (2023)
19. Vincent, J.: Google's ai-powered smart compose feature is coming to docs (2019), https://www.theverge.com/2019/11/20/20973917/google-docs-smart-compose-feature-g-suite-update
20. Whalen, J., Mouza, C., et al.: Chatgpt: Challenges, opportunities, and implications for teacher education. Contemporary Issues in Technology and Teacher Education (2023)
21. Wooldridge, M.: A brief history of artificial intelligence: what it is, where we are, and where we are going. Flatiron Books (2021)
22. Yuan, A., Coenen, A., Reif, E., Ippolito, D.: Wordcraft: story writing with large language models. In: ACM IUI (2022)