

Exploring User Interface Constraints for Reading and Writing

by

Nikhita Joshi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirements for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2025

© Nikhita Joshi 2025

EXAMINING COMMITTEE MEMBERSHIP

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Parmit Chilana Associate Professor, School of Computing Science, Simon Fraser University
Supervisor	Daniel Vogel Professor, Cheriton School of Computer Science, University of Waterloo
Internal Members	Géry Casiez Professor, Department of Computer Science, Université de Lille Adjunct Professor, Cheriton School of Computer Science, University of Waterloo
	Edith Law Associate Professor, Cheriton School of Computer Science, University of Waterloo
Internal-external Member	Oliver Schneider Associate Professor, Department of Management Science and Engineering, University of Waterloo

AUTHOR'S DECLARATION

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

STATEMENT OF CONTRIBUTIONS

This dissertation includes first-authored peer-reviewed material that has appeared in conference and journal proceedings published by the Association for Computing Machinery (ACM). The ACM's policy on reuse of published materials in a dissertation is as follows:¹

"Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included. Authors can use any portion of their own work in presentations and in the classroom (and no fee is expected)."

The following list serves as a declaration of the works included in this dissertation. This material is expanded and revised from the original publication:

Portions of Abstract, Chapter 1, and Chapter 2:

Nikhita Joshi. 2023. User Interface Constraints to Influence User Behaviour when Reading and Writing. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23 Adjunct)*. <https://doi.org/10.1145/3586182.3616710>

Portions of Chapter 3:

Nikhita Joshi, Géry Casiez, and Daniel Vogel. 2025. Is Pagination Better than Scrolling when Reading on a Phone? In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. <https://doi.org/10.1145/3706599.3720178>

Portions of Chapter 4:

Nikhita Joshi and Daniel Vogel. 2024. Constrained Highlighting in a Document Reader Can Improve Reading Comprehension. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. <https://doi.org/10.1145/3613904.3642314>

Portions of Chapter 5:

Nikhita Joshi and Daniel Vogel. 2025. Writing with AI Lowers Psychological Ownership, but Longer Prompts Can Help. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*. <https://doi.org/10.1145/3719160.3736608>

¹ <https://authors.acm.org/author-resources/author-rights>

ABSTRACT

Constraints are fundamental to human-centred design. Although by definition, constraints “limit” or “restrict” the capability of software, when designed correctly, they can have enabling characteristics as well. I sought to understand how user interface constraints can positively affect user outcomes, in ways that go beyond traditional goals of error-proofing.

Drawing from different theories in psychology, this dissertation presents four projects that study different types of user interface constraints while reading and writing. First, in a passive reading context, I evaluate the effects of two commonly-used document navigation techniques on reading comprehension: scrolling, where the reader has complete control over the viewport position; and pagination, where the viewport is restricted to specific locations. Second, in an active reading context, I propose limiting the number of words that can be highlighted in document reader software to improve reading comprehension scores. Third, when writing with large language models, I propose requiring users to write longer prompts so they feel more psychological ownership towards the generated output, and design interaction techniques to nudge users to write longer prompts. Finally, when prompting large language models to learn more about documents, I propose restricting where prompting can occur by requiring it to be anchored to specific text in the document and design new commenting techniques with different requirements that must be satisfied to finalize the comment.

Overall, this dissertation demonstrates the potential of user interfaces and interaction techniques that purposely constrain users. Through several controlled experiments, my findings suggest that user interface constraints may not be effective for certain activities, like when passively reading, but when actively reading and writing with large language models, they can often encourage positive user outcomes, like improved reading comprehension and psychological ownership. However, designing interaction techniques that leverage ‘soft’ constraints is challenging, and such interaction techniques do not always nudge users. Together, this dissertation contributes knowledge on the effectiveness of non-error-proofing constraints when reading and writing, and interaction techniques that can be integrated into current reading and writing interfaces.

ACKNOWLEDGMENTS

Daniel Vogel, I will forever be grateful, honoured, and proud to have been your student. You taught me everything I know about HCI research and none of this would have been possible without you. Thank you for believing in me and my research, for guiding me when I got lost, and for being so kind throughout. Choosing you as my advisor was the easiest decision of my life. Words cannot even express just how much I will miss you.

Thank you to my amazing committee, Parmit Chilana, Oliver Schneider, Géry Casiez, and Edith Law, for your feedback and for making my defence such a positive and memorable experience. I owe special thanks to Géry and Edith for their feedback on my research proposal a couple of years ago; my dissertation became a lot more focused and cohesive thanks to you two.

I am lucky to have collaborated with so many wonderful people, but I owe special thanks to a few in particular. Géry Casiez, merci beaucoup pour tes conseils sur les chapitres 3 et 4, et pour ta générosité pendant mes visites à Lille. Antony Albert Raj Irudayaraj, Allen Yeh, and Matthew Lakier, I had a blast working with you, and I am so happy to call you three my friends. Thank you to Nathalie Henry Riche, Michael Glueck, Christopher Collins, and Parastoo Abtahi for mentoring me during research internships. I also thank Fraser Anderson for his continued mentorship and for broadening my exposure to HCI research in ways that directly inspired this dissertation.

To everyone in the HCI Lab, thank you for being such wonderful lab mates and friends. I have cherished our daily coffee walks, gossiping over lunch, our board game nights and celebratory dinners, pushing for deadlines together, and the constant good vibes. Special thanks to the original squad; Johann, Greg, and Lisa; my baobeis; Cheryl, Alê, Margaret, Delara, Felicia, Lucy, Linda, and Selena; the aspiring professors' support group; Damien, Allen, and Matthew; and my other Exii brothers; Ryan, Antony, and Brandon.

To my friends, thank you for the entertaining and chaotic group chats, weekend getaways, and the endless memes, laughter, and support. Special thanks to the Millidgeville girls; Jenny and Margaret; my MTA friends; Maria, Émilie, and Olivier; my spicy ranters; Cheryl and Rebecca; my St. Paul's pals; Iara, Cheryl, and Chaitanya; my Meta cruwu; Cheryl, Chad, David, and Geonsun; and Matt.

Last, but certainly not least, I thank my family for their unwavering love and support. Mummya and Shiva, thank you for the countless WhatsApp calls and for being my #1 cheerleaders. Janaki and Brian, thank you for letting me borrow your nice and expensive clothes, for letting me use your streaming subscriptions so I could watch trashy reality TV shows, and for always being there when I needed a big sister and brother to lean on. Thank you, Grey, for being the best and cutest playmate. Most of all, thank you, Mom and Dad, for encouraging me to pursue graduate studies, for instilling a strong work ethic in me, for celebrating my highs and cheering me up during my lows, and for always prioritizing my well-being and happiness. I love you all so much.

Thank you all for having such a profound impact on my life.

TABLE OF CONTENTS

	List of Figures	ix
	List of Tables	xi
	1 INTRODUCTION	1
	1.1 Research Goals and Objectives	3
	1.2 Contributions	5
	1.3 Dissertation Outline	7
	2 RELATED WORK	8
	2.1 Constraints in Human-Centred Design	8
	2.2 Duality of Constraints	9
	2.3 Friction in Design	10
	2.4 Constrained Reading and Writing	11
	2.5 Summary	11
	3 EVALUATING THE EFFECT OF DOCUMENT NAVIGATION TECHNIQUES ON READING COMPREHENSION	13
- Navigating -	3.1 Motivation	13
	3.2 Background and Related Work	14
	3.3 Experiment	16
	3.4 Discussion	20
	3.5 Conclusion	22
	4 CONSTRAINING HIGHLIGHTS IN DOCUMENT READER SOFTWARE TO IMPROVE READING COMPREHENSION	23
— Highlighting —	4.1 Motivation	23
	4.2 Background and Related Work	24
	4.3 Experiment 1: Constrained Highlights	25
	4.4 Experiment 1: Results	30
	4.5 Discouraging Excessive Highlighting	35
	4.6 Experiment 2: Interaction Techniques	39
	4.7 Discussion	40
	4.8 Conclusion	43
	5 REQUIRING LONGER PROMPTS TO INCREASE PSYCHOLOGICAL OWNER- SHIP WHEN WRITING WITH AI.	44
- Prompting -	5.1 Motivation	44
	5.2 Background and Related Work	46
	5.3 Experiment 1: Prompt Length	47
	5.4 Experiment 1: Results	51
	5.5 Experiment 2: Plateau	55
	5.6 Encouraging Longer Prompts in a Real Interface	57

5.7	Experiment 3: Delayed Submission	59
5.8	Experiment 3: Results	63
5.9	Experiment 4: Pumping a Slider	66
5.10	Receiving Suggestions from AI	72
5.11	Experiment 5: Prompt Suggestions	72
5.12	Discussion	77
5.13	Conclusion	81
6	INTEGRATING PROMPTING WITH DOCUMENT TEXT THROUGH AI MARGIN NOTES	82
6.1	Motivation	82
6.2	Background and Related Work	83
6.3	AI Margin Notes	86
6.4	Experimental Method	88
6.5	Experiment 1: Integration	91
6.6	Experiment 2: Selection Automation	94
6.7	AI Margin Note Techniques	98
6.8	Experiment 3: Human and AI Involvement	100
6.9	Discussion	104
6.10	Conclusion and Future Work	106
7	CONCLUSION AND FUTURE DIRECTIONS	108
7.1	Summary of Contributions	108
7.2	Effectiveness of Constraints for Reading and Writing	109
7.3	Difficulties with Designing New Interaction Techniques	110
7.4	Other Possibilities for Constrained Reading and Writing	111
7.5	Beyond Reading and Writing	112
7.6	Who Should Impose Constraints?	113
7.7	Final Word	113
	BIBLIOGRAPHY	114
	Glossary	134

LIST OF FIGURES

Figure 1.1	Research path showing specific chapters, research questions, methodologies, and main results.	4
Figure 3.1	<i>Reading Comprehension</i> for SCROLL and PAGE.	18
Figure 3.2	<i>Reading Duration</i> for SCROLL and PAGE.	19
Figure 3.3	Task workload metrics for SCROLL and PAGE.	19
Figure 4.1	Constrained highlighting reading interface.	27
Figure 4.2	Individual and average <i>Total Words Highlighted</i> for the first pilot study.	28
Figure 4.3	Individual and average <i>Reading Comprehension</i> by condition for the second pilot study.	28
Figure 4.4	Individual and average <i>Reading Comprehension</i> by condition for the third pilot study.	29
Figure 4.5	<i>Reading Comprehension</i> by HIGHLIGHTS.	31
Figure 4.6	<i>Total Words Highlighted</i> by HIGHLIGHTS.	31
Figure 4.7	Example highlighted content.	32
Figure 4.8	Example differences between similar highlights.	33
Figure 4.9	Distribution of highlight start locations.	33
Figure 4.10	Timeline showing when the limit was reached.	33
Figure 4.11	Distribution of all highlighting attempts for CON- STRAINED.	34
Figure 4.12	<i>Reading Duration</i> by HIGHLIGHTS.	34
Figure 4.13	Questionnaire scores by condition after the reading portion of the experiment.	35
Figure 4.14	Questionnaire scores by condition after the test portion of the experiment.	35
Figure 4.15	Interaction techniques that make the act of highlighting more difficult.	36
Figure 5.1	Writing a short story with a minimal prompt and a detailed prompt.	45
Figure 5.2	Experimental interface.	49
Figure 5.3	<i>Psychological Ownership</i> by CONDITION.	52
Figure 5.4	<i>Duration</i> by CONDITION.	52
Figure 5.5	<i>Word Count</i> by CONDITION.	53
Figure 5.6	<i>Text Similarity</i> by CONDITION.	53
Figure 5.7	<i>Mental Demand and Effort</i> by CONDITION.	54
Figure 5.8	<i>Psychological Ownership</i> by CONDITION.	57
Figure 5.9	Initial prompt entry interface designs.	58
Figure 5.10	Initial prompt response interface designs.	58
Figure 5.11	<i>Final Word Count</i> by CONDITION for the initial pilot experiment with all interface modifications.	59
Figure 5.12	Experimental interface with a time delay.	61
Figure 5.13	<i>Final Word Count</i> by CONDITION.	64
Figure 5.14	Prompt word count each time the submit button was pressed.	65

Figure 5.15	Augmented prompt entry techniques.	67
Figure 5.16	<i>Word Count</i> by CONDITION.	70
Figure 5.17	<i>Psychological Ownership</i> by CONDITION.	71
Figure 5.18	<i>Mental Demand, Physical Demand, Frustration, and Effort</i> by CONDITION.	71
Figure 5.19	<i>Word Count</i> by CONDITION when receiving AI-generated suggestions.	75
Figure 5.20	<i>Psychological Ownership</i> by CONDITION when receiving AI-generated suggestions.	75
Figure 5.21	<i>Mental Demand, Physical Demand, Frustration, and Effort</i> by CONDITION when receiving AI-generated suggestions.	76
Figure 6.1	Different ways to leverage LLM capabilities in document reader software.	83
Figure 6.2	Integration design parameter.	86
Figure 6.3	Selection automation design parameter.	87
Figure 6.4	Human and AI involvement design parameter.	88
Figure 6.5	Experimental reading interface.	90
Figure 6.6	Techniques tested in Experiment 1.	93
Figure 6.7	Techniques tested in Experiment 2.	96
Figure 6.8	Experiment 2 results.	97
Figure 6.9	Techniques tested in Experiment 3.	98
Figure 6.10	Experiment 3 results.	103

LIST OF TABLES

Table 1.1	Design space exploring <i>User Activity</i> and <i>Constraint Type</i> .	2
Table 3.1	Demographic information for participants.	16
Table 3.2	Statistical test results for all metrics.	18
Table 4.1	Participant demographics for Experiment 1.	26
Table 4.2	Statistical test results for all metrics.	30
Table 4.3	Participant demographics for Experiment 2.	39
Table 5.1	Demographic information for Experiment 1.	48
Table 5.2	Omnibus and post hoc statistical tests for Experiment 1.	51
Table 5.3	Example prompt types submitted by participants.	54
Table 5.4	Demographic information for Experiment 2.	55
Table 5.5	Omnibus and post hoc statistical tests for Experiment 2.	56
Table 5.6	Demographic information for Experiment 3.	60
Table 5.7	Omnibus and post hoc statistical tests for Experiment 3.	63
Table 5.8	Demographic information for Experiment 4.	66
Table 5.9	Omnibus and post hoc statistical tests for Experiment 4.	69
Table 5.10	Demographic information for Experiment 5.	73
Table 5.11	Omnibus and post hoc statistical tests for Experiment 5.	74
Table 5.12	Between-subjects statistical tests comparing NONE and SUGGESTIONS.	77
Table 6.1	Demographics for Experiment 1.	92
Table 6.2	Statistical test results for Experiment 1.	93
Table 6.3	Demographics for Experiment 2.	95
Table 6.4	Statistical test results for Experiment 2.	97
Table 6.5	Demographics for Experiment 3.	100
Table 6.6	Omnibus statistical test results for Experiment 3.	101
Table 6.7	Post hoc statistical test results for Experiment 3.	102
Table 7.1	Expanded version of the design space explored in this dissertation.	109

INTRODUCTION

If you are hiking in the woods and come across a cliff with a sharp drop, you would immediately stop walking. If you need to open a door without a handle or knob, your instinct would be to push it open. If given the choice between napping in an armchair or on a sofa, you would opt for the sofa.

Such decisions are fundamental to everyday life, but what exactly drives them? According to ecological psychology, humans interact with the environment by perceiving what can and cannot be done within it as they try to achieve their goals [63]. Cues that tell us what we cannot do *constrain* our actions. More formally, *constraints* are defined as mechanisms that limit or restrict something, such as “our ability to do something” [121, p. 6]. More specifically, the sharp drop at the edge of the cliff hinders the ability to walk, the lack of a handle or door knob prevents the act of pulling the door, and the width of the armchair is inconducive to lying down. In other words, physical properties of the environment constrained specific actions.

Now, suppose you are completing a form on a website. You try to click the “Submit” button, but it is disabled, preventing you from clicking it. Upon seeing this, you verify the contents of the form and realize that you forgot to complete one field. After entering the missing information, the “Submit” button then becomes enabled, allowing you to click it. What caused you to review the information you had entered in the form and edit your responses accordingly? Much like we observed with the previous examples, the user interface provided cues for what could and could not be achieved within it, and the cues that indicated restricted actions formed *user interface constraints*. Specifically, the disabled “Submit” button prevented you from submitting an incomplete form.

Constraints like these are a fundamental part of user interface design as they “guide the actions” of users [128, p. 67]. However, I argue that user interface constraints have primarily been used to prevent *errors*, which are things that prevent the system from working correctly or prevent someone from using the system correctly. Some examples of such error-proofing constraints include preventing users from typing letters into an input field that is intended for phone numbers, and disabling the “Undo” command when no command history exists.

As designers, I believe that we should pay closer attention to the power of user interface constraints. Though by definition, constraints “limit” and “restrict,” prior work outside of user interface design suggests that constraints have several enabling properties too. For example, working under a short time limit can increase productivity and focus [43], and restrictions on thought can lead to creative breakthroughs and innovation [26, 163, 167].

Inspired by this enabling quality of constraints, user interface designers have been increasingly constraining user interactions in non-error-proofing ways. For example, time restrictions can regulate smartphone use [93, 94] and encourage more focused knowledge-sharing [84, 112], character limits

“The power of constraints has largely been ignored.”
– Don Norman [129]

		Constraint Type	
		Where	When
User Activity	Reading	CHAPTER 3	CHAPTER 4
	Writing		CHAPTER 5
	Reading + Writing	CHAPTER 6	

Table 1.1: Design space exploring *User Activity* and *Constraint Type*. I keep referring to this design space as chapters are introduced.

can encourage more user participation on social media [81], and displaying only a few search results can improve user confidence and satisfaction when retrieving information in a search engine [135]. Even some more ubiquitous user interface constraints have such enabling properties: requiring a password to be a certain length enables increased account security [157], and restricting the placement of objects to pre-defined locations in vector graphics software enables more precise and aesthetically pleasing designs [14]. These types of constraints are not about reducing errors; rather, they exist to encourage important and positive outcomes for the people being subjected to them and the tasks people are trying to achieve. This is not to say that preventing errors is not a positive outcome, but these types of constraints really restrict *when* and *where* interaction can occur, putting them into the forefront of the user’s minds during interaction.

Reading and writing are important activities that can foster cognitive skills like the ability to self-reflect, think critically, and pay attention [87, 183]. Moreover, we read and write daily through user interfaces: we read articles on the go by scrolling on our phones, highlight and write embedded comments in document reader software while studying, and now with the advent of large language models (LLMs), we even read and write documents by writing prompts for generative AI assistants. These interactions may be simple, but they can also introduce problems for users. Scrolling can make it difficult for readers to simultaneously process and relate new text to text that was previously seen [141, 154]. Readers often have difficulty identifying important information from documents, causing them over-highlight [10], and when taking notes, many tend to take verbatim notes [28], which can hinder comprehension [10, 27]. Readers can use LLMs to summarize specific parts of documents [98], but these interactions occur in separate chat-based interfaces, which requires more effort to refer to specific text in the document [118] and risks disrupting reading due to additional cognitive load [73, 115]. Writers who use LLMs may feel like the generated output is no longer ‘theirs,’ which may reduce their motivation to contribute, especially as AI capabilities become increasingly integrated into user interfaces [36].

My focus on reading and writing was also a self-imposed constraint, which enabled me to think of more research ideas.

Could these issues faced while reading and writing be mitigated through user interface constraints?

To explore the potential of user interface constraints in the context of reading and writing, I conducted four projects that explore a design space of *Constraint Type* and *User Activity* (Table 1.1). Specifically, I explore *Constraint Types* that restrict *where* an interaction can occur or an interface element can

be placed, and *when* an interaction can occur; and *User Activities* that focus on *reading*, *writing*, and combinations of *reading and writing*. My research goal is to better understand whether different types of user interface constraints can improve user outcomes for different kinds of user activities. A secondary goal of mine is to design novel interaction techniques that leverage the power of constraints, but in ways that give users more flexibility. Specifically, through ‘soft’ constraints that increase the cost of interaction without preventing it.

1.1 RESEARCH GOALS AND OBJECTIVES

To summarize, the goal of this dissertation is to:

Evaluate the effects of non-error-proofing user interface constraints while reading and writing, and design interaction techniques that use ‘soft’ constraints.

To achieve this objective, we answer the following research questions, as shown in the research path (Figure 1.1):

- (a) *Do user interface constraints that restrict where the viewport is placed benefit **passive reading** on a phone?* To answer this question, we conducted a large-scale study to compare two commonly-used document navigation techniques for reading comprehension. Our results suggest that constraining the placement of the viewport through pagination does not significantly improve reading comprehension.
- (b) *How can user interface constraints that restrict when interaction can occur benefit highlighting, a common **active reading** strategy?* To answer this question, we conducted a large-scale study that compares different word limits for text highlighting. Our results suggest that constraining how many words readers can highlight can improve reading comprehension.
- (c) *Given the promising results of (b), we were led to the follow-up question: can interaction techniques that make the act of highlighting more difficult encourage readers to highlight fewer words?* To answer this question, we designed three new interaction techniques that add different types of costs as readers highlight more words, and conducted exploratory pilot experiments to evaluate their effectiveness. Our results suggest that these interaction techniques either did not reduce the number of highlighted words, or did not improve reading comprehension.
- (d) *How can user interface constraints that restrict when interaction can occur benefit writing with a generative AI assistant?* To answer this question, we conducted two experiments that required writers to write prompts that had varying word length requirements. Our results suggest that requiring longer prompts can improve feelings of **psychological ownership**, feelings of the writing ‘belonging’ to the writer [139].
- (e) *Given the promising results of (d), we then sought to answer: can interaction techniques that make the act of submitting a short prompt more difficult encourage*

I use “we” when discussing specific projects to acknowledge the contributions of my collaborators.

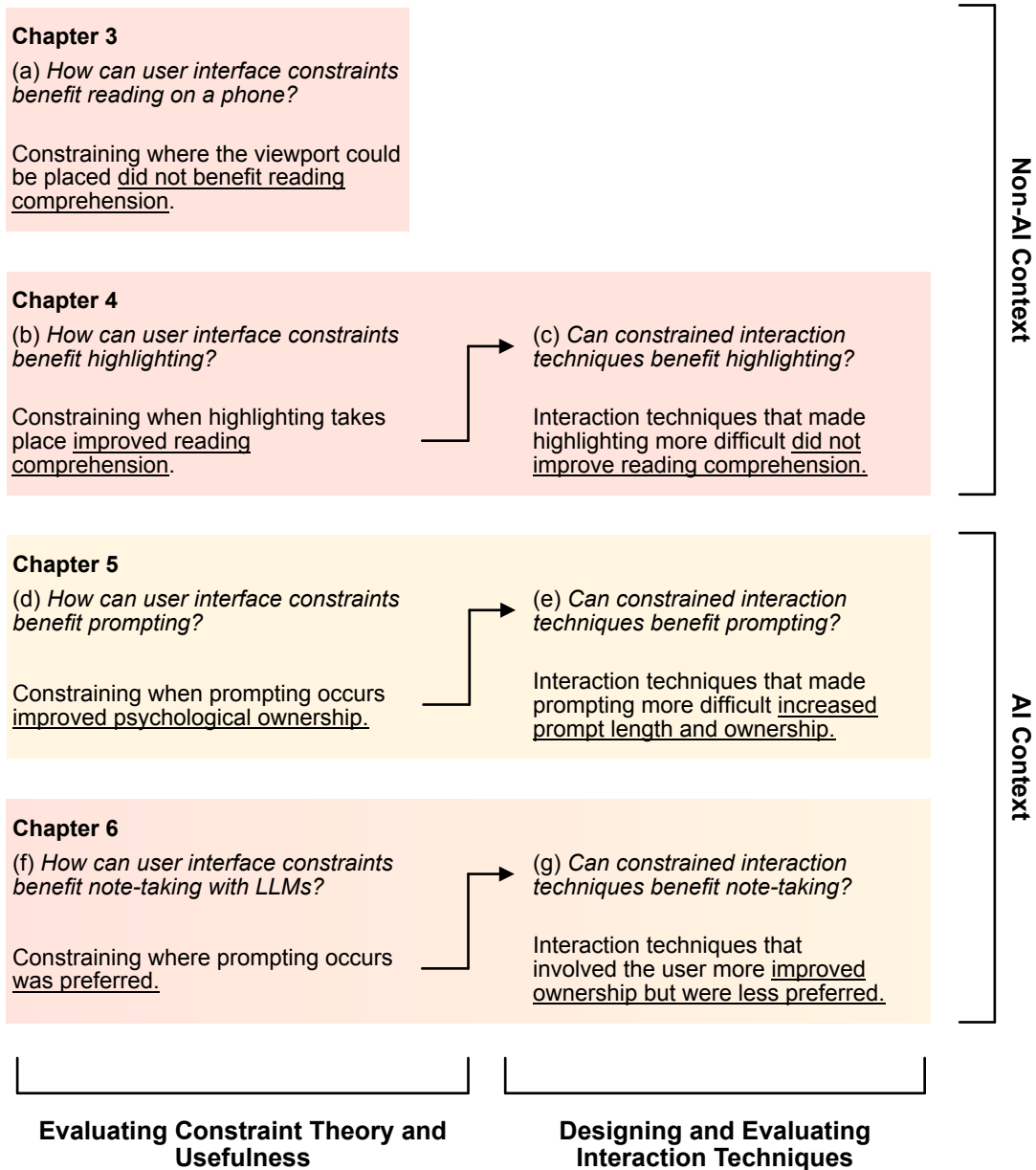


Figure 1.1: Research path showing specific chapters, research questions and dependencies (arrows), and main results (underlined). Orange indicates reading projects and yellow indicates writing projects (Chapter 6 has both reading and writing). Note that this does not reflect the order in which the projects were completed, and any discrepancies that arise from this will be discussed in margin notes.

users to write longer prompts? To answer this question, we designed and evaluated new interface modifications and conducted multiple user studies to evaluate their effectiveness at increasing prompt length and psychological ownership. Our results suggest that techniques that increased temporal and motor costs were effective at increasing prompt length and improving psychological ownership.

- (f) *How can user interface constraints that restrict where prompting can take place benefit note-taking, a common active reading strategy?* To answer this question, we propose “AI margin notes” that integrate prompting into the commenting feature of document reader software by requiring users to position prompts and responses relative to text in the document. Our results suggest that this increased integration was preferred and valued.
- (g) Building on the technique developed to answer (f), we expanded our investigation to the question: *how can user interface constraints that restrict when notes can be saved benefit note-taking?* To answer this, we designed and developed several AI margin note techniques that involve the user and the LLM to varying degrees. Several techniques required users to meet certain requirements before they could continue. Our results suggest that techniques with requirements generally required more mental demand and effort than receiving summaries, and were associated with more psychological ownership.

1.2 CONTRIBUTIONS

Four projects explore these research questions. Each contains at least one user study that tests the effectiveness of the *Constraint Type* for the *User Activity*. All user studies were conducted online, with diverse groups of participants. If the constraint was found to be effective, we then designed and evaluated novel interaction techniques that tried to elicit the same behaviour as the constraint.

1.2.1 Comparing Document Navigation Techniques when Passive Reading

Looking at a common type of constrained interaction, we compare two document navigation techniques when reading on a smartphone: scrolling and pagination. When scrolling, readers have full control over where the viewport is positioned, but this is restricted to pre-defined locations with pagination. We conducted a large-scale ($n=100$), between-subjects online experiment where participants read a short story using either pagination or scrolling, and answered multiple-choice questions. Our results found no significant differences between these two techniques for reading comprehension, duration, and task workload, which differs from prior work. This suggests that constraining the position of the viewport while reading with pagination may not benefit readers.

1.2.2 Constraining the Number of Words Readers Can Highlight

We investigate if an imposed user interface constraint limiting the number of highlighted words in a document reader can help readers be more selective when highlighting text and improve reading comprehension. We conducted a large-scale ($n=127$), between-subjects online experiment where participants read a short story and could highlight nothing, up to 150 words, or an unlimited number of words. Our results suggest that constraining the number of words that can be highlighted is associated with higher reading comprehen-

sion scores than highlighting nothing or highlighting an unlimited number of words. We design and evaluate new interaction techniques that make the act of highlighting more temporally, physically, and perceptually demanding, yet found these techniques did not change behaviour or improve reading comprehension.

1.2.3 *Requiring Longer Prompts when Writing with LLMs*

We report on two experiments that examine the relationship between psychological ownership and prompt length. Participants wrote short stories either completely by themselves or wrote prompts of varying lengths. Results show that when participants wrote longer prompts, they had higher levels of psychological ownership. Their comments suggest they thought more about their prompts, often adding more details about the plot. However, benefits plateaued when prompt length was at least three quarters of the target story length. To encourage users to write longer prompts, we evaluated two interaction techniques that modify the prompt entry interface of chat-based generative AI assistants: pressing and holding the prompt submission button, and continuously moving a slider up and down when submitting a short prompt. Our results suggest that these techniques increased prompt length and were associated with higher psychological ownership than baseline techniques. We further augmented these techniques by showing AI-generated suggestions for how the prompts could be expanded. This further increased prompt length, but did not lead to improvements in psychological ownership. Our results show that simple interface modifications like these can elicit more writing from users and improve psychological ownership.

1.2.4 *AI Margin Notes for Prompting and Note-Taking*

We explore integrating AI into document comments, a concept we formalize as “AI margin notes.” Three design parameters characterize this approach: margin notes are integrated with the text while chat interfaces are not; selecting text for a margin note can be automated through AI or manual; and the generation of a margin note can involve AI to various degrees. We conducted three within-subjects experiments to evaluate this approach. In each experiment, participants read short stories and interacted with one prompting or note-taking technique before completing reading comprehension tests. Two experiments investigate integration and selection automation, with results showing participants prefer integrated AI margin notes and manual selection. We then designed six alternative techniques that involve the human and the AI in different ways. A third experiment revealed that techniques with less AI involvement resulted in more psychological ownership, but faster and less effortful designs are generally preferred. Surprisingly, the degree of AI involvement had no measurable effect on reading comprehension. Our work shows that AI margin notes are desirable and contributes implications for their design.

1.3 DISSERTATION OUTLINE

The rest of this dissertation is structured as follows:

- [Chapter 2](#) discusses relevant related work, including how constraints are used in human-centred design, the duality of constraints, friction in design, and previous work that has constrained reading and writing.
- [Chapter 3](#) studies scrolling and pagination document navigation techniques and the effects they have on reading comprehension, reading duration, and task workload.
- [Chapter 4](#) proposes constraining the number of words readers can highlight in document reader software and examines the effects this has on reading comprehension.
- [Chapter 5](#) explores the relationship between prompt length and psychological ownership and proposes requiring users to write longer prompts for increased psychological ownership.
- [Chapter 6](#) proposes AI margin notes that integrate prompting with document text when using LLMs to read documents and write notes by requiring users to first select text that anchors prompts and responses. It then presents techniques that require users to fulfill specific requirements to save their AI margin notes.
- [Chapter 7](#) summarizes our contributions, discusses implications for design, and presents opportunities for future work.

RELATED WORK

In [Chapter 1](#), I defined constraints as mechanisms that limit or restrict something. They are fundamental to almost every domain, though they may take on other names such as “requirements,” “conditions,” and “rules.” Each domain uses different types of constraints. For example, musicians need to consider subjective or stylistic constraints [[114](#), [136](#)], but businesses need to consider input, output, and process constraints [[1](#)]. Different domains are also associated with different levels of constraint: the arts, for example, are typically considered to be less constrained than something like engineering [[166](#)]. In this chapter, we discuss how constraints have been used in human-centred design, the duality of constraints as both an enabler and inhibitor, friction in design, and previous work that has constrained reading and writing.

2.1 CONSTRAINTS IN HUMAN-CENTRED DESIGN

Constraints are fundamental to human-centred design as they “guide the actions” of users [[128](#), p. 67]. But what is the basis of this? Ecological psychology suggests that properties of the environment indicate actions that can and cannot be performed within it, a concept that was formalized by Gibson when he defines *affordances* as “what it offers the animal, what it provides or furnishes, either for good or ill” [[63](#), p. 119]. One example he discusses is how large, rigid, and roughly “knee-high” surfaces afford sitting. However, for a shorter child, a surface that is “knee-high” for an adult would not afford sitting. In his work, Gibson does not formally define “constraints,” but this description of affordances implicitly suggests that properties of the environment can restrict what can be achieved within it.

Inspired by Gibson, Norman makes the role of constraints more explicit in design. Notably, he argues that people will *perceive* different affordances from the environment, and that such “perceived affordances” are more important than all possible “affordances.” For example, people will perceive a chair to afford sitting, even though it can also afford standing.¹

Norman notes how designers need to “ensure that the desired, relevant actions are readily perceivable” [[129](#)]. One way this can be achieved is by using constraints to guide users towards using something as intended. He described four types of constraints [[128](#), p. 125-132], primarily based on physical objects, and later linked them to user interface design [[129](#)]. First, physical constraints are those that rely on physical properties, like constraining the location of a cursor to the screen’s coordinates. Second, cultural constraints restrict design choices to follow norms or conventions. For example, visualizing a scroll bar on the right of the screen follows left-to-right reading patterns of Western

“Different layouts afford different behaviours for different animals, and different mechanical encounters.” – James J. Gibson [[63](#), p. 120]

¹ Originally, Norman did not distinguish between “affordances” and “perceived affordances,” but later clarified that he actually meant “perceived affordances” [[129](#)] as he realized (through conversations with Gibson) that affordances can exist without necessarily being perceived. This change was made in revised versions of “The Design of Everyday Things.”

cultures, and using commonly-understood keyboard shortcuts like “Ctrl+Z” for “Undo” follows design conventions. Third, logical constraints rely on human logic and inductive reasoning, almost acting as constraints on the user’s thought. For example, if an interface references a button that is not currently visible on the screen, the user will likely deduce that they need to scroll down to see it. Finally, semantic constraints restrict design choices based on knowledge of the situation and the world, like preventing letters from being typed into a field for a phone number.

This classification provides clarity on the types of constraints that may exist in a user interface, however, I believe that constraints are primarily incorporated into user interfaces to reduce user and system errors. Restricting the position of the cursor ensures that ‘click’ events are properly mapped to user interface elements; using norms and conventions for commands reduces the chances of the user issuing incorrect commands learned from other software; using inductive reasoning and logic to scroll reduces risk of the user abandoning the system after perceiving an error that does not exist; and ensuring that only numbers can be entered into a phone number field prevents data validation and processing errors.

2.2 DUALITY OF CONSTRAINTS

In other domains, constraints are often used as measures to turn large or ill-defined problems into those that are smaller and well-defined [146, 161]. This can “make certain thoughts – certain mental structures – possible” [26, p. 58].

These benefits have long-been understood in creative domains, with some even claiming that constraints are a fundamental part of the creative process. Elster [58, p. 176] describes the creative process as two-staged: the choice of constraints and choosing within constraints. Similarly, Stokes [167] argues that creative breakthrough can be achieved through self-imposed constraints that are manipulated throughout the creative process. Self-imposed time limits can spark creative thinking and solutions [53, 76, 151] and reduce perfectionism [43]. Prior work in HCI has shown the success of constrained creativity systems that purposely limit the range of actions or choices users have for the visual arts [11, 168] and music [13, 45].

Constraints can be useful beyond creativity applications as well. When applied to search engines, limiting the number of results improves a user’s confidence and satisfaction in their search [135], and having more requirements to satisfy when searching for information can result in more in-depth and relevant searches [18]. Social media applications with character limits like Twitter can lower the user’s perceived effort for generating each post, encouraging more participation [81]. Time-bounded video-sharing social media like TikTok forces content creators to focus on one or two key points in each video, which can be more comprehensive and enjoyable for viewers [112]. Self-imposed time restrictions and cumbersome unlocking mechanisms can be effective for discouraging excessive smartphone usage [93, 94]. During my master’s, I created MicroMentor [84], a help-seeking system that constrains help sessions to three minutes. Results showed that constrained help sessions

can encourage more help-seeking, as there is less of a perceived burden associated with asking for and giving help.

However, there is a balance that needs to be struck as over constraining can have adverse effects. Over-constrained creativity experiences can result in too much uniformity and less enjoyment [48, 88, 89], and severe time constraints may cause stress [76, 109]. Many researchers have proposed that there is an inverted U-shaped relationship between the level of constraint and the extent to which constraints are enabling [6, 58]. This has been described as “the sweet spot” [134, 151]. Optimal results can be achieved by balancing the level and type of constraint [1].

2.3 FRICTION IN DESIGN

This means there must also be a balance when designing constrained interaction techniques: over-constraining could deter users from interacting with the system, yet under-constraining likely does not benefit users. This could be achieved by identifying optimal levels of constraint for each interaction and for every type of user. But perhaps an easier and more practical approach to give users more flexibility is making constraints ‘softer.’ Unlike ‘hard’ constraints that *prohibit* interactions, ‘soft’ constraints *modify* interactions to guide users towards a similar outcome. This can be done by making interactions slower or more difficult without preventing them.

Though it may feel counterintuitive, prior work shows that this can be beneficial. According to [dual process theory](#), humans have two ways of thinking: “System 1” thinking that is fast, automatic, and intuitive; and “System 2” thinking that is slow, deliberate, and reflective [85]. Friction in design can encourage users to move from System 1 to System 2 thinking that is more conducive to increased attention and analytical thinking [30, 32, 46, 85]. This can effectively [nudge](#) users by encouraging positive changes in behaviour without preventing user actions [32, 46].²

Friction can be incorporated into user interfaces through constraints. For example, Wang et al. [180] showed that time delays that prevented social media post submissions until enough time had passed can encourage increased reflection before posting to social media. Lewis and Vogel [106] used time delays to encourage users to transition to expert-mode menu selection techniques. When using an AI assistant, Buçinca et al. [30] required users to wait before displaying AI-generated suggestions, which encouraged more analytical thinking about the output and reduced over-reliance [30]. Beyond time delays, other types of friction have also been successful. For example, Soboczenski et al. [164] purposely obscured text in a transcription task and this was associated with fewer errors. Similarly, Cockburn et al.’s “frost-brush” technique [44] purposely obscured interface widgets, but allowed users to ‘brush’ them off by moving their cursor over the widgets repeatedly. This approach improved spatial tasks in interfaces, like learning the location of keys on a virtual keyboard.

² Thaler and Sunstein’s original definition of a nudge was specifically for interventions that are also “easy and cheap to avoid” [169, p. 6]. However, many nudges in HCI arguably do not have this property, like when they are used to introduce friction [32].

2.4 CONSTRAINED READING AND WRITING

Some prior work has explored the impacts of constraints on reading and writing. When reading, Walczyk et al. [174] examined the effects time limits have on reading comprehension scores and found that reading under mild time pressure encouraged readers to stay focused and motivated while reading, which resulted in better scores. Similarly, Chang [38] found that when learning a second language, imposing time limits on reading can improve reading comprehension as readers are forced to read faster and focus on broader concepts. To encourage such benefits, some prior work has proposed constrained reading techniques to encourage reading at a certain pace or in a more focused way. For example, Breznitz and Share [29] showed that students can recall more information from text if it is ‘erased’ after reading it. Though not strictly enforced, some websites like Medium³ show estimated read times for each article and some marketing research suggests they can improve user engagement [162]. However, not all techniques have been successful, for example, rapidly presenting document text one word at a time can hinder comprehension and increase fatigue [12]. Auto-scrolling features that prevent users from controlling the pace of reading can help users read faster, but they may hinder comprehension [156] and cause fatigue at higher speeds [77].

When writing, prior work has also explored constraints that control the pace of writing. Biskjaer et al. [19] explored how time constraints that are strictly enforced in a custom writing tool can encourage more creative writing. They found writing under time constraints generated more new writing, but the quality was lower. However, such writing would be suitable as a warm-up exercise, or to help writers overcome perfectionism. Others have explored including more specifications through writing prompts and inspirational materials. Haught and Johnson-Laird explored how different writing prompts affect creative writing [74]. They tested using nouns and drawings, which are more limiting as they impose more specifications on what needs to be written, and found that drawings resulted in longer and more creative sentences. Neate et al. [123] develop a system, MakeWrite, for writing erasure poetry, a constrained form poetry that is created by removing words from an existing piece. MakeWrite allows users to control two constraints related to the creation of erasure poetry: the amount of text retained and the number of words that can be grouped together from the original piece. Results from a user study showed that the system was effective at enabling people with speech and language impairments to express their creativity, something participants claimed would have been difficult to achieve otherwise.

2.5 SUMMARY

Overall, previous work shows the importance of constraints in human-centred design, but has mostly applied it to reduce errors. From other domains, we know that constraints can have other positive effects, which could benefit reading and writing more specifically, but this has not been as well-studied. When reading, prior work has primarily focused on using time limits and

³ <https://medium.com>

visualization techniques to encourage reading at a certain pace. When writing, prior work has also explored the role of time limits, but has primarily focused on evaluating the benefits of constraining thought through inspirational materials and prompts to spark creative writing. My dissertation builds on this prior work by evaluating the effectiveness of constraints and designing interaction techniques for a wider range of user activities, like highlighting and writing with LLMs, and using a wider range of constraints such as word limits.

EVALUATING THE EFFECT OF DOCUMENT NAVIGATION TECHNIQUES ON READING COMPREHENSION

In this chapter, we compare two commonly-used document navigation techniques for [passive reading](#) on a phone: scrolling, where the reader has an unconstrained ability to control the viewport position; and pagination, where the user cannot control *where* the viewport is positioned. We explore the effects on reading comprehension, duration, and task workload.

3.1 MOTIVATION

The fundamental interactions to navigate documents longer than a single device screen are *pagination* and *scrolling*. Pagination separates a document into discrete pre-defined chunks (“pages”) mimicking the structured experience of reading a physical document. Usually, a single tap on a button or screen location moves to the next or previous page. Scrolling uses the metaphor of sliding the document under a screen-sized viewport (or the complementary metaphor of sliding a viewport over the document). A vertical drag on the touch screen moves the document up or down (or a scrollbar is used to move the viewport position in the document).

Clearly, these two techniques use very different interaction mechanics, and that changes how people read a document. Readers have complete control over how they view a document when scrolling since they decide how much the document advances, for example, by a few paragraphs or a few lines. But readers do not have such control with pagination, as the position of the document (or viewport) is fixed to specific locations. In other words, pagination is a spatially constrained document navigation technique, but scrolling is spatially unconstrained.

Research has found that having an unconstrained ability to navigate document text through scrolling is beneficial for certain reading goals, like skimming a document [154]. However, several studies suggest that the structure imposed by pagination leads to better reading comprehension [61, 141, 154]. The theory is that the discretization of a document into pre-defined pages makes it easier for readers to build “a better mental representation of the text,” which subsequently makes it easier for readers to relocate information and remember details [141]. Perhaps for this reason, popular smartphone reading apps like iOS Books use pagination.

Yet, all of the studies above are more than 10 years old or were conducted on desktop computers. Today, smartphones are truly ubiquitous and so is scrolling, not just for many reading experiences like online articles, but for many other types of interactions like viewing social media posts. Unlike desktop systems, touchscreen scrolling on a phone uses direct manipulation of the relative document position, which has highly desirable instrumental interaction properties: no spatial indirection, high compatibility, and high integration [7]. These interaction mechanics and the prolific use of smart-

	Where	When
R	CH. 3	CH. 4
W		CH. 5
R+W	CH. 6	

*This project explores constraining **where** the viewport can be placed while **reading** a document.*

phones could mean that people have become better at reading by scrolling and therefore better at remembering what they read. The benefits associated with the imposed structure of pagination may no longer apply. A 2022 study that tested smartphones claims some “marginal” overall benefit to pagination [75], but this may be lacking in ecological validity as all participants were recruited from education programs at a university and participants had to use a provided device rather than their own device. We argue that there is a need to revisit this fundamental question in the context of modern smartphone users to answer the research question: *does pagination still lead to better reading comprehension than scrolling?*

We conducted a highly-controlled yet ecologically valid between-subjects experiment where 100 participants, who were recruited from the general public using a crowdsourcing service, were given a short story to read on their own smartphone using only scrolling or pagination. We recorded how long they spent reading, examined their reading comprehension using a standard objective time-limited test, and collected responses to NASA-TLX-inspired questions. Our findings found no significant differences between scrolling and pagination, which is an important contribution as it challenges a long-held assumption about the benefits of different reading techniques.

3.2 BACKGROUND AND RELATED WORK

The differences between scrolling and pagination for reading comprehension have long been studied in desktop environments. In 2009, Sanchez and Wiley [154] were the first to find statistically significant differences between scrolling and pagination. In a between-subjects experiment, 40 participants read informational text on a desktop by scrolling or paginating and then wrote an essay to demonstrate reading comprehension. Their results showed that pagination had statistically higher essay scores ($p < 0.01$). One possible explanation offered by the authors is *working memory capacity*: “the ability to process and store information simultaneously.” Specifically, readers with a lower working memory capacity may not be able to simultaneously process and reveal new information when scrolling and relate it to existing relationships in the text. In fact, their results showed that participants with a lower working memory capacity were more negatively impacted by scrolling than those with a higher working memory capacity ($p < 0.05$).

A related concept to working memory capacity is mental workload. In 2008, Wästlund et al. [185] conducted two between-subjects experiments on a desktop with 40 participants each where participants read a document (the type of document was not specified) and simultaneously responded to dialog boxes that appeared on the screen. The time taken to close these pop-ups was used to measure mental workload. After reading, participants answered 4-6 multiple-choice questions to measure reading comprehension. Their results showed that scrolling induces more mental workload than pagination ($p = 0.02$ for Experiment 1; $p = 0.002$ for Experiment 2), but they did not find any differences in reading comprehension ($p = 0.79$ for Experiment 1; $p = 0.38$ for Experiment 2). However, they note a possible floor effect due to the low number of questions.

In 1997, Piolat et al. [141] conducted a between-subjects experiment, where 26 participants read informational text on a desktop by scrolling or paginating and wrote summaries of it. They claimed that participants who navigated the document with pagination tended to receive higher grades ($p = 0.18$) and included more details in their summaries ($p = 0.06$) due to marginal statistical significance (i.e., $0.05 < p < 0.1$). They did not observe any significant differences nor any trends for reading duration.

Since these initial explorations on desktops, more recent work has compared scrolling and pagination on smartphones, yet they also only report marginal significance. Fukaya et al. [61] conducted a within-subjects experiment in 2011, where 12 participants read narrative and procedural texts by paginating or scrolling on a smartphone. The two text types were used to examine different metrics. Reading comprehension was evaluated using narrative texts, and participants answered fill-in-the-blank questions. Pagination had marginally significant improvements in reading comprehension (exact p -values not given). “Operation time” was evaluated using the procedural texts. Specifically, these texts contained instructions for a virtual console shown on a desktop, so participants read the text while simultaneously interacting with the virtual console. Overall, there were no significant differences in operation time, but participants who preferred scrolling were marginally faster than those who preferred paginating ($0.05 < p < 0.1$). Task workload for reading the procedural test was quantified using a NASA-TLX, and they found that scrolling had lower mental load than paginating, which contrasts Wästlund et al.’s findings [185].

A more recent 2022 between-subjects study by Haverkamp et al. [75] examined the effect of both device size and technique on text understanding. 145 participants read a scientific document by scrolling or paginating on a smartphone or tablet (32-39 participants per condition) and before answering a long-answer question about the document. They conclude that overall, pagination tends to result in better understanding of the text due to a marginally significant main effect of technique ($p = 0.07$).

To summarize, prior work examining both desktops and smartphones largely suggests that pagination may lead to improved reading comprehension over scrolling. However, most studies, including both studies that used smartphones, merely show ‘trends,’ rather than conclusive and significant statistical findings. As such, the validity of these claims is unclear. Furthermore, with the exception of Haverkamp et al.’s work [75], all of these studies are over 10 years old, and people may have become better at reading by scrolling due to several years of additional practice scrolling on smartphones. Even the recent study by Haverkamp et al. may be lacking in ecological validity as all participants were recruited from education programs at a university and participants did not use their own personal devices that they would be most comfortable with. For these reasons, we believe it is important to replicate these experiments.

Table 3.1: Demographic information for participants.

Gender		Age		Education	Handedness	
Men	58	18-24	6	Less than High School	3	Right 87
Women	36	25-34	42	High School Diploma	13	Left 11
Non-Binary	4	35-44	35	Some University (no credit)	19	Ambidextrous 2
Unknown	2	45-54	10	Technical School	8	
		55-64	3	Bachelor’s Degree	46	
		65-74	4	Professional Degree Beyond Bachelor’s Degree	3	
				Master’s Degree	6	
				Doctorate	2	

3.3 EXPERIMENT

The goal of this experiment is to examine the effect of document navigation techniques on reading comprehension. Participants read short stories and answered questions about it, before answering questions about their overall experience. Our hypothesis was that pagination would result in higher reading comprehension scores, given the promising results from prior work.

3.3.1 Participants

Prior work mostly recruited participants from universities, specifically, psychology courses. To increase the diversity of participants, we recruited 114 participants through the online crowdsourcing platform, Prolific.¹ The task was restricted to the United States and Canada, those who had completed at least 1,000 tasks on the platform, and those whose approval rating was greater than 98%. Crowdsourced experiments have an increased risk of fraudulent responses, so we manually examined all open-ended responses to identify fraudulent responses (i.e., very short responses like “nice” and “good”) [153], but only 1 participant was omitted for this reason. We also removed 13 participants who experienced technical difficulties with our interface. In total, 14 were omitted from analyses, leaving 100 valid responses (Table 3.1). All self-reported being proficient at reading in English (all ≥ 5 on a 1-7 point scale). Participants reported spending 4 hours on their phones (IQR = 3.25), and 1 reading on their phones (IQR = 1) daily. Upon completing the experiment, participants were remunerated \$7.50. For each condition, participants who scored within the top 25% for reading comprehension received an additional \$3 bonus as additional incentive to do well on the test.

I use mostly the same inclusion and exclusion criteria in all chapters. I fine-tuned these parameters during an internship, where I conducted crowdsourced experiments using Amazon Mechanical Turk (MTurk). The data quality for Prolific is generally better than MTurk [52], so looking back, I probably did not have to make the inclusion criteria as strict.

3.3.2 Reading Comprehension Task

Participants read one of ten short stories ($\sim 1,500$ words each) from easyCBM [4]. easyCBM is a system developed by the University of Oregon that provides teachers with standardized benchmark assessments for a variety of subjects.² We selected stories from the 8th grade reading level and used the accompany-

¹ <https://www.prolific.com>

² We contacted the researchers via email and received special permission to use these stories.

ing 20-question multiple-choice tests to evaluate reading comprehension. An 8th grade reading level is representative of documents and websites targeted for the general public [126].

3.3.3 Apparatus

The experimental software was a React web application. The reading interface displayed the short story and was styled to resemble a Medium article. Participants could navigate through the story by scrolling (i.e., they used their finger to move the text) or by paginating. Pagination was designed to mimic the mechanics of reading on a Kindle: participants could touch the left and right of the screen to move to the previous and next page respectively, and a thin 5 pixel progress bar at the top of the screen indicated where they were in the document.

The test interface displayed 20 multiple-choice questions, which were navigated through by scrolling. A toolbar at the top of the screen indicated how many questions the participant had answered and the remaining time.

3.3.4 Procedure

Participants received a link to the experimental software through the Prolific system. The task was restricted to smartphones. First, they answered questions about demographic information and read instructions. Second, they performed a basic system calibration, where they rotated their phones to be in landscape and portrait mode and the system automatically captured device information, screen size, and logged a photo of how the story was rendered on the device for debugging purposes. For participants assigned to the pagination condition, the calibration phase was also used to split the story into pages that were adapted to the device screen size and orientation. Upon completing this calibration, they read the story. There was no time limit for the reading phase, but after reading the story, they completed the multiple-choice reading comprehension test, which had a 5 minute time limit. This duration was selected for the test after a pilot study, where we observed that participants took 7.7 minutes on average to complete the test, so 5 minutes felt appropriate to increase test difficulty. Participants could finish the test early if they answered all questions, otherwise, the test would automatically end after 5 minutes. Finally, they answered questions about their overall experience.

3.3.5 Design

This is a between-subjects experiment with one independent variable, **TECHNIQUE**, with two levels: **SCROLL** ($n=51$) and **PAGE** ($n=49$). All participants read one of 10 short stories. The story and **TECHNIQUE** were both randomly assigned. The primary measures computed from logs were *Reading Comprehension*, representing the number of questions correctly answered on the reading comprehension test (0-20 range); *Reading Duration*, the time taken to read the story in minutes; and *Test Duration*, the time taken to complete the test in minutes. We also collected subjective measures about task workload using

This pilot study was conducted for the highlighting project discussed in Chapter 4. Therefore, the protocol varies slightly, but we discuss it in more depth in the following chapter.

Table 3.2: Statistical test results for all metrics.

Metric	<i>U</i>	<i>p</i>	<i>RBC</i>
<i>Reading Comprehension</i>	1262	0.93	0.01
<i>Reading Duration</i>	1022	0.12	0.18
<i>Test Duration</i>	1094	0.28	0.12
<i>Mental Demand</i>	1137	0.43	0.09
<i>Physical Demand</i>	1349	0.39	0.08
<i>Temporal Demand</i>	1273	0.87	0.01
<i>Effort</i>	1111	0.34	0.11
<i>Performance</i>	1279	0.83	0.02
<i>Frustration</i>	1186	0.61	0.05

similar wording as the NASA-TLX: *Mental Demand*, *Physical Demand*, *Temporal Demand*, *Effort*, *Performance*, and *Frustration*. All were interval data on a 1-7 numerical scale, and *Performance* was reversed (i.e., $8 - x$) to align numeric scores and valence.

3.3.6 Results

For all metrics, we analyzed the distribution and the homogeneity of variance of the data using a Shapiro-Wilk test and a Levene’s test. While the data for SCROLL and PAGE had similar variances, none of the metrics followed a normal distribution. Therefore, we use Mann-Whitney U tests in the analysis to follow. However, as a quick test, we also tried using independent samples t-tests as they are generally robust to non-normal distributions with large sample sizes and have more statistical power, yet we received similar results. Table 3.2 shows detailed statistical test results.³

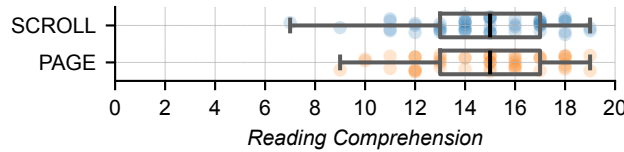


Figure 3.1: Reading Comprehension for SCROLL and PAGE.

3.3.6.1 Reading Comprehension

As demonstrated in Figure 3.1, we did not observe any significant differences in Reading Comprehension between SCROLL and PAGE (MDN = 15, IQR = 4 for both). This finding differs from prior work [61, 75, 141, 154], but aligns with Wästlund et al.’s findings [185].

³ RBC is the rank-biserial correlation and is a measure of effect size for rank-based statistical tests.

3.3.6.2 Duration

Similarly, as shown in Figure 3.2, we did not observe any significant differences between SCROLL (MDN = 6.2, IQR = 4.2) and PAGE (MDN = 7.7, IQR = 4.8) for *Reading Duration*. This finding differs from Fukaya et al. [61], who found that scrolling was marginally faster than pagination, but aligns with Piolat et al. [141] who did not find any significant differences.

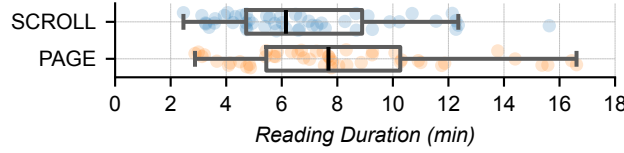


Figure 3.2: Reading Duration for SCROLL and PAGE.

For *Test Duration*, we did not find any significant differences between SCROLL (MDN = 3.2, IQR = 1.3) and PAGE (MDN = 3.6, IQR = 1.4).

3.3.6.3 Task Workload

We did not observe any significant differences between SCROLL and PAGE for any task workload metrics (Figure 3.3). Notably, there were no significant differences between SCROLL (MDN = 3, IQR = 3) and PAGE (MDN = 4, IQR = 2) for *Mental Demand*, which differs from Wästlund et al. [185], who found that scrolling techniques had higher mental workload than pagination, and from Fukaya et al. [61] who found that scrolling is associated with lower mental workload.

3.3.6.4 Screen Rotation, Zoom, and Pan

For increased ecological validity, participants could scroll in either landscape or portrait mode and could also zoom and pan across the screen. As an additional check to ensure that these interactions did not hinder internal validity, we also analyzed the above metrics using participants who only read in portrait mode and who did not zoom or pan. 26 were briefly omitted ($n=42$ for SCROLL; $n=32$ for PAGE), with most being from the PAGE condition. Even then, we did not observe any significant differences between SCROLL and PAGE.

Reflecting on these results after seeing a talk given by G  ry Casiez, I suspect this may have been due to the between-subjects design [149]. Participants did not have an ‘anchor’ to make relative comparisons. Chapter 5 and Chapter 6 use within-subjects designs.

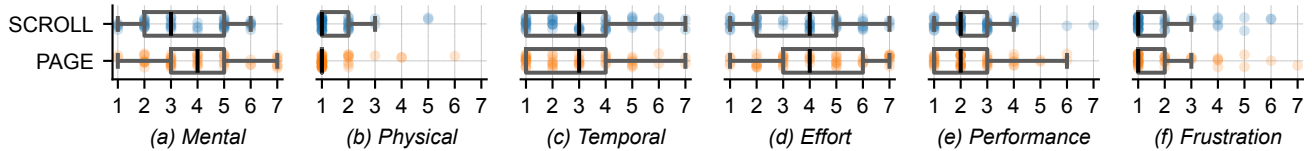


Figure 3.3: Task workload metrics for SCROLL and PAGE: (a) *Mental Demand*, (b) *Physical Demand*, (c) *Temporal Demand*, (d) *Effort*, (e) *Performance*, (f) *Frustration*.

3.3.6.5 Open-Ended Responses

We examined all free-form responses to better understand user experiences. As expected, most participants did not perceive any issues or differences when scrolling (e.g., *“being able to scroll while reading was effortless”* (P100)). However, P36 wished they *“did not have to scroll so much”* while reading and how this was cumbersome: *“I found myself wanting to read more of the story at once, without having to stop and scroll and re-find my position that I was reading.”* This idea of navigation being difficult while scrolling has been demonstrated in prior work [141].

Pagination led to more varied opinions. Several participants noted how the design was similar to that of a Kindle or e-reader, so they did not experience any issues. Others enjoyed having the story split into multiple pages, as it was *“easier to read”* (P54), provided *“a sense of progression”* (P34), made *“the process less intimidating”* (P12) and *“the story more digestible”* (P43), and helped them *“focus even more”* (P32). Two participants noted how pagination made them *“forget things [like they were] different stories”* (P2) and that pagination *“made it even more difficult to keep track of everything”* (P5). These concerns may be related to working memory capacity, however, prior work suggests that this is typically associated with scrolling rather than paginating [154].

3.4 DISCUSSION

To summarize, our results did not find any significant differences in reading comprehension between scrolling and pagination when reading on a smartphone, which differs from findings presented in prior work. Participants took roughly the same amount of time to read the stories in both conditions, and there were no significant differences in perceived task workload.

These results differ from prior work, which has largely suggested that scrolling results in worse reading comprehension [61, 75, 141, 154]. We suspect that the ubiquity of smartphones, and subsequently scrolling, over the past several years has made it easier for people to create mental representations of the text they are reading, which is important for reading comprehension [141, 154].

Although we did not find significant differences between scrolling and paginating, the open-ended responses suggest that pagination strategies can still be desirable for readers and possibly improve things like focus and a sense of progression. As such, giving users the option to switch between pagination and scrolling modes when reading on a smartphone may further improve user experience.

3.4.1 Limitations

There are several experimental design factors that may have influenced the results. First, participants were reading with a clear goal in mind, which may have encouraged deep reading. However, there are other types of reading tasks, such as skimming, or reading without a particular goal, like when browsing through articles online. It is possible that scrolling and pagination may lead to more pronounced effects for different reading types.

Similarly, there are several different types of documents that are commonly read on a smartphone. We chose to focus on short stories at an 8th grade reading level as this does not require any background knowledge from participants and it represents the target reading level of websites [126]. Prior work primarily focused on non-fiction and scientific documents [75, 141, 154], so replicating this with the same types of documents may reveal differences between the two techniques.

Another factor is when the test took place. We administered the reading comprehension test immediately after the participants read the story. This, coupled with the document type, may have led to higher scores on the test. Repeating this experiment with a break in between the reading and testing phases may allow us to more accurately determine whether these effects were determined by the reading technique or if they were a result of short-term memory. However, it should be noted that prior work also administered both phases within the same experimental session.

The nature of the test may have an effect as well. Wästund et al. [185] were the only others to find no differences in reading comprehension between scrolling and pagination and they also administered multiple-choice questions. Others relied on written responses [75, 141, 154] or fill-in-the-blank questions [61], so it is possible that scrolling and pagination may help readers do better on certain types of questions, which warrants additional research.

We estimated the power of our sample sizes using the standard deviation of all scores ($SD = 2.6$). We found that we could have detected medium to large effect sizes ($d = 0.57$), corresponding to ~ 1.5 point differences in scores (7.4%), which is slightly less than one letter grade of a difference. As such, we believe that our sample size is large enough to detect what we believe to be meaningful differences between the two conditions if they exist. However, it is still possible that smaller differences exist between the two techniques, which only would have become apparent with additional statistical power from a larger sample size.

Finally, Sanchez and Wiley [154] suggested that people with lower working memory capacity may be more negatively impacted by scrolling than those who have high working memory capacity. To keep the experiment duration short, we did not collect information about working memory capacity, however, we did recruit from a diverse population. Controlling for working memory capacity may reveal some differences between the two techniques.

3.4.2 Future Work

An important direction for future work is to investigate the effects of long-term use and to explore other metrics beyond reading comprehension. For example, the act of scrolling can cause readers to enter a flow state, which tempts them to keep scrolling even while they are engaged with and enjoying the displayed content. This can result in fragmented reading, which can be dangerous as it may impact how people interpret and understand media like the news [66]. Similarly, pagination could possibly reduce “mindless scrolling” tendencies on social media. It is possible that pagination may mitigate such risks in the long-term, but more work is needed to study this.

Chapter 4 uses the same short stories, but Chapter 6 uses non-fiction documents.

Subsequent studies have longer breaks between reading and test phases: 24 hour breaks in Chapter 4, and 2 hour breaks in Chapter 6.

We did not conduct an a priori power analysis as we were confident in our ability to detect significant differences given prior work and the results we obtained in Chapter 4. However, we were surprised by these results, which is why we decided to conduct a post hoc power analysis.

Our participants suggested that pagination helped improve focus, and studying this in more depth and across a wider variety of tasks would be beneficial. For example, in an information retrieval context, Kim et al. [95] found that participants found more relevant search results when paginating than when scrolling on a smartphone. To enable such investigations, an important technical contribution would be an app or browser extension that can turn any continuous scrolling webpage into a paginated version.

3.5 CONCLUSION

We investigate the effects of paginating and scrolling when reading on a smartphone. Our results indicated no significant differences between these two techniques for reading comprehension, duration, and task workload, which differs from prior work that has studied these effects on desktop and mobile devices. We believe that our findings are important as they challenge a long-held assumption about the pros and cons of different reading techniques, and they do so using a task with high ecological validity.

CONSTRAINING HIGHLIGHTS IN DOCUMENT READER SOFTWARE TO IMPROVE READING COMPREHENSION

In the previous chapter, we compared the effects of pagination, which constrains *where* the viewport can be placed, to scrolling, which gives readers complete control over its position. Although existing theory suggested that the pagination document navigation technique would improve reading comprehension scores, we did not find differences between scrolling and pagination. However, there are other theories in psychology that suggest that constraints may be beneficial in during *active reading*. In this chapter, we propose limiting how many words readers can highlight in document reader software, limiting *when* highlighting can occur, to improve reading comprehension scores.

4.1 MOTIVATION

Marking up existing text with underlines and highlights (“text-marking” [10]) is a common technique used by readers to remember information from documents [60]. Prior work in learning theory suggests this is caused by two main effects. First, marking up text visually isolates it from other text, making it more memorable [127]. Second, by considering whether some text is important and worth highlighting, readers think more about about it, resulting in better recollection [47, 192]. However, to reap these benefits, readers must be selective by only marking what is truly important [60, 127]. Yet many tend to over-mark text [10], such as when using text-marking to help concentration while reading [127]. This is problematic since it can create less visual separation between important and unimportant text, hindering recollection, and it can instill a false sense of comprehension [23, 192]. Effective text-marking strategies can be taught. For example, Leutner et al. [105] helped readers to reflect on their highlighting through self-regulation training, but this required lots of time and effort as readers had to follow a lengthy 90-minute training program consisting of almost 50 slides.

To force readers to highlight fewer words, document reader software could impose a constraint on text-marking. This is easy to do, but the effect of such a limit on comprehension has not been examined. Constraining how much text can be marked within a document reader should implicitly encourage readers to reflect on their markings and self-regulate. Specifically, the reader must reconsider whether each marking is truly important, and revise accordingly if the limit has been reached to regain the ability to mark new text. This process forces the reader to play a more active role with text-marking and should lead to a highlighted document with adequate visual separation.

We conducted a large-scale between-subjects experiment (n=127) in which participants were assigned a level of text highlighting constraint: no highlights, up to 150 highlighted words, and unconstrained highlights. The results show that participants subjected to a highlight constraint performed better in a reading comprehension test taken 24 hours later. To encourage readers to

	Where	When
R	CH. 3	CH. 4
W		CH. 5
R+W	CH. 6	

*This project explores constraining **when** highlighting can occur while **reading** a document.*

highlight fewer words without a strict word limit, we created and evaluated new interaction techniques that introduce temporal, motor, and perception costs when highlighting text. These techniques were either ineffective at reducing the number of words readers highlight, or reductions in the number of highlighted words did not transfer to improved reading comprehension. Although such ‘soft’ constraints were ineffective, our work contributes the **first exploration of user interface constraints in the context of text-marking**, and it is the first to show that **constrained highlighting can improve reading comprehension scores** without traditional self-regulation training. More broadly, our work validates theories in psychology, and applying this theory within HCI could lead to several research directions for the design of document reader interfaces.

4.2 BACKGROUND AND RELATED WORK

Our work is related to prior work on text-marking strategies in psychology and prior work that has imposed constraints or augmented highlights in text editors.

4.2.1 *Benefits of Text-Marking*

The reason why text-marking can be an effective **active reading** strategy has been debated in psychology: is it because of the *act* or the end *result*? According to **levels of processing theory** [47], information is recalled for longer periods of time the deeper the information has been processed, which can be achieved through “a greater degree of semantic or cognitive analysis,” like making associations to prior knowledge and experiences. Yue et al. [192] gave students a passage to read and they were instructed to study for a test that took place one week later by either just reading the passage, or reading and highlighting the passage. Their results showed that students that highlighted less received higher scores than those who were considered heavy highlighters. Yue et al. speculated that being selective while highlighting required more mental effort to decide what to highlight, which led to higher scores.

The von Restorff effect [176] states that when presented with multiple items that are similar, either visually or semantically, items that differ are more likely to be remembered. This theory can explain why text-marking can help people recall information in documents. Nist and Hogebe [127] gave students text passages with different types of information (i.e., important and unimportant details from the text) already underlined to examine whether text marks are beneficial for the resulting document they produce. Their results showed that when students are given passages with one type of information already underlined, they answered more questions about the underlined content correctly than students who received passages with another type of information highlighted.

Constraining text highlights could be beneficial for both the act and the result. Readers are encouraged to think more critically about what they highlight, which also results in a document with adequate visual separation.

4.2.2 *Pitfalls of Text-Marking*

For text-marking to be most effective, the reader should be able to distinguish between important and unimportant material [60]. Bell and Limber [10] explored the impact of reading skill on text-marking efficiency by examining the textbooks used by students in an introductory psychology course. Students with lower reading skills tended to over-highlight text, and highlighted more irrelevant information, which was associated with lower scores on the final exam than those with higher reading skills, who were more selective and focused when highlighting. Some readers may use text-markings for the wrong reasons, for example, as a concentration strategy while reading [127], which can lead to over-highlighting. Yue et al. [192] suggested that readers who do not know how to highlight effectively may feel a false sense of comprehension (i.e., “illusion of competence” [23]). When re-reading, these readers may skim over their highlights with little focus as they believe that the presence of a highlight means the content has already been encoded in memory.

Such pitfalls can be avoided by teaching readers how to highlight effectively. Using a computer training program, Leutner et al. [105] taught students how to highlight text using a five-step process, which included reflecting on their highlighting behaviours through self-regulation training. This encouraged students to monitor, self-evaluate, and make adjustments to their highlighting behaviours, and these students performed better in reading comprehension tests than those who just learned effective highlighting strategies. However, the training was lengthy, requiring readers to follow a slideshow consisting of almost 50 slides for roughly 90 minutes.

We show that constraining text highlights may encourage self-regulation by requiring the reader to monitor, self-evaluate, and adjust their highlights to adhere to the constraints imposed by document reader software.

4.2.3 *Applications in HCI*

To our knowledge, no prior work has investigated the impact of artificially constraining text-marking in a user interface. Some work has explored the positive impacts of short note-taking styles, like bullet journaling [34], on mindfulness and self-reflection [5, 170]. As described in Chapter 2, Biskjaer et al. [19] explored the effects of time constraints within a text editor to encourage more creative writing, and found that people wrote more when writing under a time constraint. Han et al.’s Textlets [72] turn text selections into interactive objects that can be manipulated and saved within a text editor to improve consistency when working under constraints imposed by technical documents. Although Textlets visually resemble text highlights, there were no limits on how many could be created within a document.

4.3 EXPERIMENT 1: CONSTRAINED HIGHLIGHTS

The goal of this experiment is to understand the impact of constrained highlighting on reading comprehension scores. Participants read a short story

Table 4.1: Participant demographics for Experiment 1.

Gender		Age		Education		English Language Proficiency	
Men	62	18-24	5	Less than High School	2	Full Professional	6
Women	60	25-34	28	High School	15	Native or Bilingual	121
Non-binary	3	35-44	48	Some University (no credit)	18		
Unknown	2	45-54	20	Technical Degree	14		
		55-64	16	Bachelor's Degree	57		
		65-74	6	Master's Degree	12		
		75+	2	Beyond Bachelor's (e.g., MD, JD)	4		
		Unknown	2	Doctorate	5		

Document Reader Frequency		Highlight Frequency		Highlight Usage	
Daily	16	Daily	3	Remember Concepts	84
Weekly	44	Weekly	31	When Commenting	37
Monthly	29	Monthly	15	Concentration	43
Less than Monthly	28	Less than Monthly	50	Other	2
Never	10	Never	18		

and were asked to highlight text in preparation for an open-book reading comprehension test 24 hours later. This is a between-subjects study where each participant could either highlight nothing, highlight up to 150 words, or highlight an unlimited number of words.

4.3.1 Participants

We recruited participants through Prolific. Participants were restricted to Canada and the United States and those who completed at least 2,500 tasks and with an approval rating greater than 98%. As was done in [Chapter 3](#), we manually examined all open-ended responses to identify fraudulent participant responses [153], but no participants were omitted for this reason. Participants were instructed not to use any other study tools or aids while reading the document, like taking notes in a separate document or taking a screenshot of it. An open-ended response asked participants if they used any tools or aids to filter out those who did not follow the experiment instructions, which has been done in other crowdsourced experiments (e.g., [107]). In total, we filtered out 15 participants (11%) who described using other study tools or aids, who experienced technical difficulties with our user interface, and who did not attempt to answer any questions during the reading comprehension test, leaving 127 valid responses ([Table 4.1](#)). Participants received \$15 in total. Much like [Chapter 3](#), for each condition, participants who scored within the top 25% received a \$3 bonus to provide a small incentive to do well on the test.

4.3.2 Task and Apparatus

Like [Chapter 3](#), participants read one of ten short stories from easyCBM and answered the associated reading comprehension test, which composed of 20

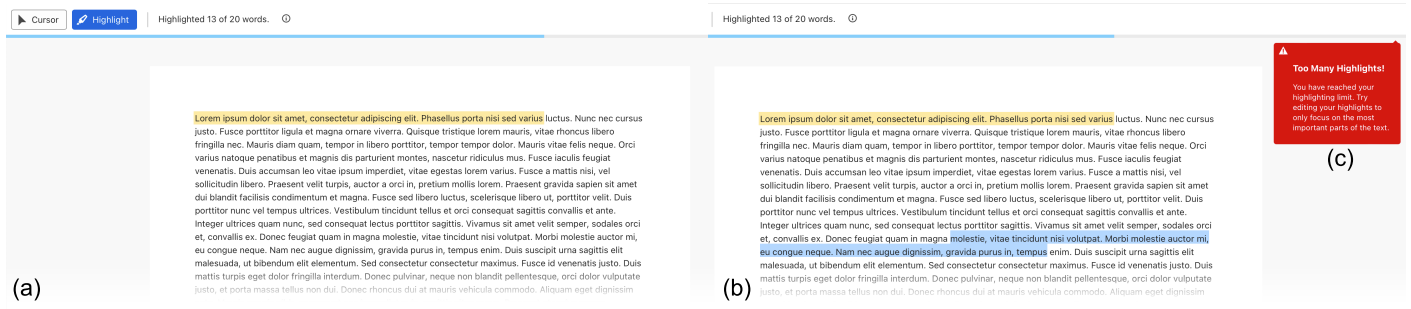


Figure 4.1: (a) Constrained highlighting interface. A count and progress bar showing how many words have been highlighted appear in the top toolbar; (b) if too many words are selected, an error message appears, and the new highlight is not created.

multiple-choice questions. A Node.js and React application (Figure 4.1) served a web-based custom document reader with interfaces for reading and testing. The reading interface displayed the document at the centre of the screen. If the participant was able to highlight, at the top was a toolbar that allowed participants to change highlighting modes, which are common in existing document readers like Adobe Acrobat. Using the “Cursor” tool, participants could first select text from the document, and then press a black “Highlight” button. Using the “Highlighter” tool, any text selected automatically became highlighted. Text selections snapped to full words. If the participant could only highlight up to 150 words, the toolbar also displayed how many words had been highlighted, and a progress bar expanded and shrunk as words were highlighted or deleted. Copying text was disabled to prevent cheating.

The test interface displayed the same document with the participant’s highlights with 20 multiple choice questions displayed on the right. The top toolbar displayed the number of questions answered, and a progress bar and countdown showed how much time remained for the test. The browser “Find” feature was disabled on the document text to prevent cheating.

4.3.3 Identifying Experimental Properties

To identify levels of constrained highlighting, we first ran a pilot experiment without any highlighting constraints with 12 participants from Prolific. This was done to better understand how many words participants naturally highlight when reading the short stories we selected for the experiment. Overall, we observed that these participants highlighted 296 words on average ($SD = 201$; Figure 4.2). Based on these results, we initially selected 250, 150, and 50 word highlight limits, which corresponds to slightly below the mean followed by decreasing 100 word intervals.

With levels of constraint identified, we then ran a pilot to understand how much time to allocate to the reading comprehension test. We ran the experiment with a 10 minute time limit for the test with 59 participants (10 to 13 per condition). We found they spent 7.7 minutes on average to complete the test ($SD = 2.0$) with a 15.1 average score ($SD = 3.1$; Figure 4.3). As the

I usually report the median and the interquartile range and show boxplots for nonparametric data. However, at this point, I had not determined the normality of the data, so I used the mean, standard deviation, and visualized 95% confidence intervals as quick checks. To better reflect how I made decisions, I report and show those here.

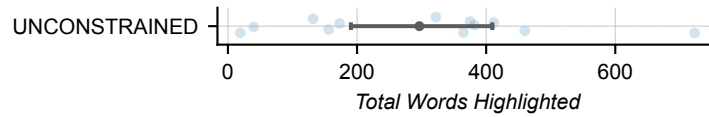


Figure 4.2: Individual and average *Total Words Highlighted* for the first pilot (n=12) to identify word limits to test in subsequent pilot studies.

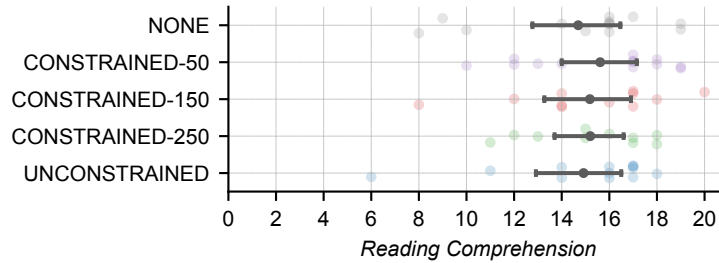


Figure 4.3: Individual and average *Reading Comprehension* by condition for the second pilot (n=59) featuring a 10 minute time limit for the reading comprehension test.

stories were short, participants were likely able to re-read them during the test, leading to higher scores and little differentiation between conditions.

These results indicated a 5 minute time limit was reasonable to increase test pressure and encourage participants to rely more on their memory and highlights, instead of re-reading the story.

As between-subjects experiments have less statistical power, it was not practical to run three experimental conditions alongside two baseline conditions at a large scale. As such, we ran another pilot with 98 participants (16 to 25 per condition) to identify which word limit was most promising. A shorter time limit proved to be successful at differentiating the conditions, and our results suggested that the 150 word limit may lead to higher scores when working under a 5 minute time limit for the test (Figure 4.4). In the main experiment, we constrain highlights to 150 words, which corresponds to roughly 10% of the document word count. We include data from these participants in the main results.

This was the pilot study I referenced in Chapter 3.

4.3.4 Procedure

Participants received a link to the document reader web application through the Prolific system. The task was restricted to desktop and laptop devices. They entered basic demographic information and read instructions, then they were presented with the reading interface where they read the short story and highlighted content of interest if permitted in their condition. There was no time limit during the reading stage of the experiment. Once they finished reading, they answered 7 short questions about their experience using the reading interface.

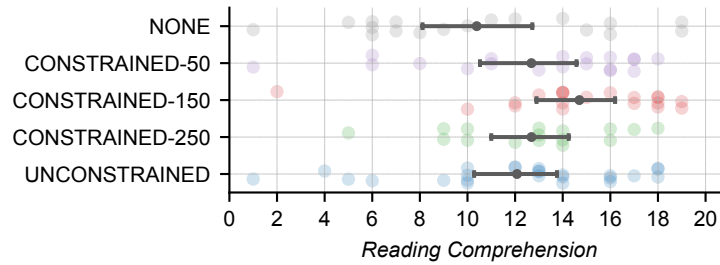


Figure 4.4: Individual and average *Reading Comprehension* by condition for the third pilot (n=98) featuring a 5 minute time limit for the reading comprehension test.

After 24 hours, the participant could access the test interface, which displayed the short story marked up with their highlights along with 20 multiple choice questions. If the participant answered all questions within the 5 minute time limit, they could press a button to finish the test early. Otherwise, the test automatically ended after 5 minutes. Participants were not allowed to pause the timer during the test, and they were told to finish the test in one sitting prior to beginning. They answered 8 short questions about their experience completing the test.

4.3.5 Design

We opted for a between-subjects design over a within-subjects design to keep the experiment shorter for each participant and to prevent order effects across conditions (e.g., implicitly learning to highlight less, or learning the types of questions that may be asked in subsequent conditions). There is one primary independent variable, **HIGHLIGHTS**, with 3 levels: **NONE** (n=43), **CONSTRAINED** (i.e., up to 150 words; n=42), and **UNCONSTRAINED** (n=42). All participants read one of 10 documents, using one **HIGHLIGHTS** condition. Both were randomly assigned.

The primary measures computed from logs were:

- *Reading Comprehension*, the number of questions the participant answered correctly during the reading comprehension test (0-20 range).
- *Words per Highlight*, the number of words within a single highlight, counting only the highlights that the participant did not delete.
- *Total Words Highlighted*, the total number of words highlighted, counting highlights that were not deleted.
- *Number of Highlights*, the final count of highlights.
- *Duration*, the time taken (in minutes) to read or highlight the document.
- *Number of Deletions*, the number of times the participant deleted a highlight from the document.

We later figured out a reasonable, within-subjects protocol to mitigate these issues, which we describe in [Section 4.6](#).

Table 4.2: Statistical test results for (a) *Reading Comprehension*, (b) *Total Words Highlighted*, (c) *Number of Highlights*, and (d) *Words per Highlight*.

comparisons		U	p	RBC	
(a) <i>Reading Comprehension</i> ($H = 15.7$, $p < .001$, $\eta_H^2 = .11$)					
NONE	CONSTRAINED	468	< .001	***	.48
NONE	UNCONSTRAINED	699	.07		.22
CONSTRAINED	UNCONSTRAINED	1146	.03	*	.30
(b) <i>Total Words Highlighted</i>					
CONSTRAINED	UNCONSTRAINED	463	< .001	***	0.48
(c) <i>Number of Highlights</i>					
CONSTRAINED	UNCONSTRAINED	910	.81		0.03
(d) <i>Words per Highlight</i>					
CONSTRAINED	UNCONSTRAINED	184465	< .001	***	0.56

- *Limit Reached*, an indicator variable for whether the word limit was exceeded while attempting to add a new highlight.

The post-reading questions had 6 measures from the NASA-TLX. The post-test questions had the same 6 measures and one additional question asking participants how frequently they referred back to the story and their highlights (all 1-7 scale). The values for *Performance* were reversed (i.e., $8 - x$) to align valence and numeric scores. The post-reading and post-test questions both included a single open-ended question asking about the participant's experience.

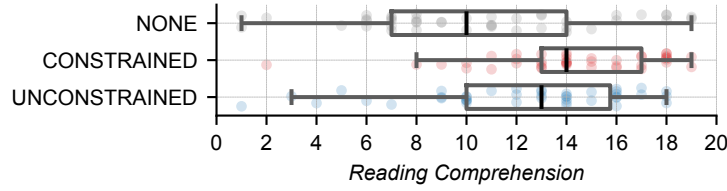
4.4 EXPERIMENT 1: RESULTS

To assess whether parametric statistical tests could be used, the assumptions of normality and homogeneity of variance were evaluated using a Shapiro-Wilk test and a Levene's test. Like we observed in [Chapter 3](#), the variances were similar, however, the data and the residuals did not follow a normal distribution. Therefore, we use a Kruskal-Wallis omnibus test and Mann-Whitney U post hoc tests with Holm's corrections for multiple comparisons. [Table 4.2](#) shows detailed statistical test results.¹

4.4.1 *Reading Comprehension*

Overall, we observe that constraining highlights to 150 words can improve reading comprehension scores ([Figure 4.5](#)). A significant effect of HIGHLIGHTS on *Reading Comprehension* and post hoc tests revealed that CONSTRAINED (MDN = 14, IQR = 4) was associated with higher scores than both NONE (MDN = 10, IQR = 7) and UNCONSTRAINED (MDN = 13, IQR = 5.75). Notably, the

¹ Although the Kruskal-Wallis is also rank-based, its effect size is typically represented using η_H^2 .

Figure 4.5: *Reading Comprehension* by HIGHLIGHTS.

spread of data for CONSTRAINED was tighter (IQR=13 to 17) than both NONE (IQR=7 to 14) and UNCONSTRAINED (IQR=10 to 15.75), suggesting that scores were more consistent when highlights were limited to 150 words.

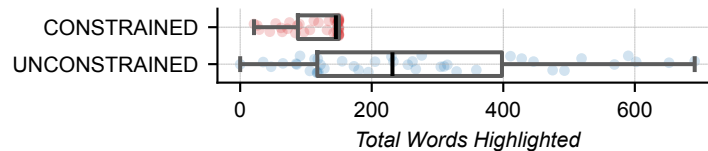
4.4.2 Highlighting Experience

To better understand why a 150 word highlight constraint improved *Reading Comprehension*, we grouped open-ended responses from the reading portion of the experiment for participants in the CONSTRAINED condition ($n=42$). I did this grouping independently, as the data was straightforward.

Eighteen participants (43%) noted that the word limit affected their highlighting strategy. Specifically, sixteen (38%) indicated that the word limit encouraged them to highlight less and focus on the the most important points, with comments like: “I kind of liked it because it forced me to highlight only the parts I thought were more important. In turn, this forced me to understand the story and main themes more” (P34).

4.4.2.1 Highlight Word Count

To corroborate these findings, we examined the *Total Words Highlighted* and *Number of Highlights* for the CONSTRAINED and UNCONSTRAINED conditions (Figure 4.6), and found that *Total Words Highlighted* was much lower in the CONSTRAINED condition (MDN = 145.5, IQR = 62.25) than the UNCONSTRAINED condition (MDN = 231.5, IQR = 280.75). However, the *Number of Highlights* for CONSTRAINED (MDN = 16.5, IQR = 15.75) and UNCONSTRAINED (MDN = 15, IQR = 11) were similar, suggesting that each highlight contained fewer words. We examined the *Words per Highlight* to confirm this and found that when CONSTRAINED, highlights were an average of 4.8 words long (MDN = 2, IQR = 5), much lower than 13.7 words in UNCONSTRAINED (MDN = 9, IQR = 14).

Figure 4.6: *Total Words Highlighted* by HIGHLIGHTS.

4.4.2.2 Highlighted Content

To get a sense of the types of words participants highlighted, we examined heat maps of the raw highlights, where common highlights between participants appeared more opaque (Figure 4.7). As we had two highlighting conditions and participants could highlight one of ten stories, the number of participants who highlighted the same story for a single condition is low (5 to 10), so we discuss common themes across all stories. Participants highlighted a wide range of text, especially for the UNCONSTRAINED condition, so we filtered the heat maps to only show text where a majority of the participants highlighted the same thing (i.e., opacity ≥ 0.5) and compared similar types of highlights across conditions. One story had no participants for the CONSTRAINED condition, so we only consider nine stories.

(a) Constrained (n=5)

This year the campers had so much to choose from! They could take swimming lessons, make candles, do leather crafts, compete in archery, make jewelry, design and paint murals on the Camp Director's van, join the 'spy craft' project and learn secret codes and make their own periscopes, or work on building the bridge across the creek. There were also cooking classes and wood burning projects to work on. Everyone was able to have swim time each afternoon, and there was a campfire led by Hi 4-H'ers at night. Then the campers went off to their own campsites to sleep in their sleeping bags until the next day when the fun would start all over again.

(b) Unconstrained (n=3)

This year the campers had so much to choose from! They could take swimming lessons, make candles, do leather crafts, compete in archery, make jewelry, design and paint murals on the Camp Director's van, join the 'spy craft' project and learn secret codes and make their own periscopes, or work on building the bridge across the creek. There were also cooking classes and wood burning projects to work on. Everyone was able to have swim time each afternoon, and there was a campfire led by Hi 4-H'ers at night. Then the campers went off to their own campsites to sleep in their sleeping bags until the next day when the fun would start all over again.

Figure 4.7: Example of highlighted content that was highlighted when (a) CONSTRAINED and (b) UNCONSTRAINED (opacity is normalized across participants).

The most common differences between similar highlights for the two conditions were removing filler words when CONSTRAINED (7 stories; 78%); in contrast, this only occurred for two stories in the UNCONSTRAINED condition (22%). This often involved separating longer highlights by article words or prepositions (Figure 4.8a). When additional adjectives were used to describe the same noun, the first adjective was typically highlighted while others were ignored (Figure 4.8b). When a person or place was described using a few sentences, participants highlighted the concept itself without the definition (Figure 4.8c). Although less information was highlighted when CONSTRAINED, some participants noted that each highlight effectively created a kind of bookmark to find additional details, for example, “it was pretty handy in order to know where to look in the text for specific segments of the story [...] Due to the limit on how much I could highlight, I only really used it for that” (P33). Some online resources at universities include highlighting tips like “highlight key words and phrases instead of full sentences” [40], so it appears as though participants in the CONSTRAINED were encouraged to highlight in this way.

4.4.2.3 Reaching the Limit

We anticipated that participants would delete more highlights in the CONSTRAINED condition once the limit was reached. Overall, there were twice

	UNCONSTRAINED	CONSTRAINED
(a)	wearing a homemade t-shirt with the word “Yes”	wearing a homemade t-shirt with the word “Yes”
(b)	Mateo was quiet, and a constant planner, researcher and designer.	Mateo was quiet, and a constant planner, researcher and designer.
(c)	Dr. Garcia had been their family doctor since even before Angie was born.	Dr. Garcia had been their family doctor since even before Angie was born.

Figure 4.8: Example differences between similar highlights when UNCONSTRAINED and CONSTRAINED.

as many *Number of Deletions* when compared to UNCONSTRAINED (117 vs. 59). However, participants only reached the word limit 46 times (3% of all highlighting activities). This suggests that participants were highlighting few words from the onset rather than shortening retroactively, supported by comments like: *“the amount of highlighted words was something that I had to constantly keep track of. I predicted that I would run out of highlights available unless I used them carefully”* (P31).

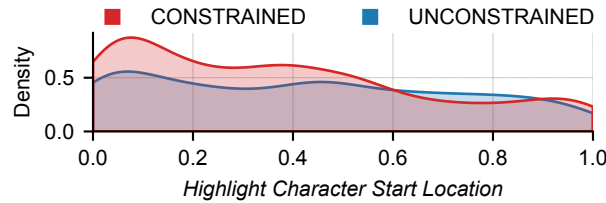


Figure 4.9: Distribution of highlight start locations by condition (normalized across documents).

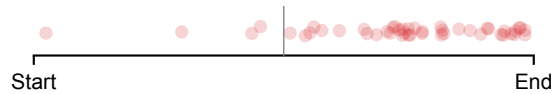


Figure 4.10: Timeline showing when participants in the CONSTRAINED condition hit the 150 word limit (normalized across participants).

This reluctance to delete highlights is further supported by the distribution of highlight locations for all valid highlights (Figure 4.9). Participants in the CONSTRAINED condition tended to highlight more at the beginning of the document, while those in the UNCONSTRAINED condition highlighted more consistently throughout the entire document. By focusing their highlights on text earlier in the document, participants typically ran out of words halfway through their reading and highlighting session (Figure 4.10), which led to more invalid highlighting attempts for text later in the document (Figure 4.11).

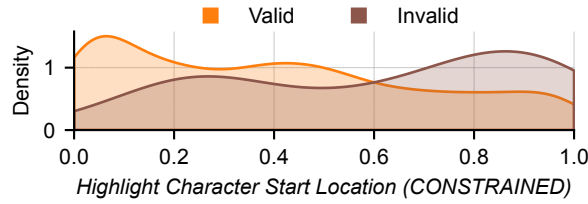


Figure 4.11: Distribution of valid and invalid highlighting attempts for the CONSTRAINED condition (normalized across documents).

One participant noted that running out of words at important moments of the story, which typically occurred halfway through the documents we selected, was frustrating: *“it was slightly frustrating that I ran out of words right when I got to the climax of the story”* (P75).

4.4.2.4 Duration

Although participants in the CONSTRAINED condition had to adopt new strategies and think more about highlighting under a word limit, they took roughly the same amount of time (MDN = 9.6, IQR = 7.2) as those in the UNCONSTRAINED condition (MDN = 8.3, IQR = 4.9; Figure 4.12). We anticipated that NONE would be faster than both CONSTRAINED and UNCONSTRAINED, but no significant differences were observed (MDN = 9, IQR = 6.9).

4.4.3 Subjective Feedback

Overall, all conditions were rated similarly for all metrics post reading (Figure 4.13) and post test (Figure 4.14). All median scores were equal to or below a “neutral” score of 4 post reading, but all conditions had scores greater than a “neutral” score for *Mental Demand*, *Temporal Demand*, and *EFFORT* post test. Although eleven participants from the CONSTRAINED condition (26%) had indicated feelings of frustration or increased mental demand in open-ended responses after reading and highlighting the story, this did not seem to impact scores.

Like I noted in Chapter 3, we may not have observed significant differences due to the between-subjects design.

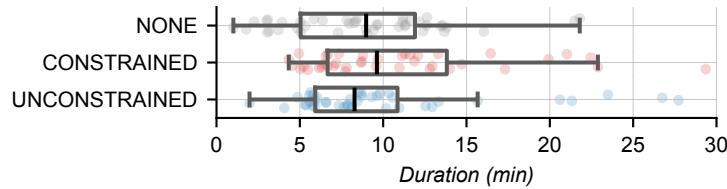


Figure 4.12: Reading *Duration* by HIGHLIGHTS. Note that 3 points with values greater than 30 minutes are not shown to improve visibility of the box plots.

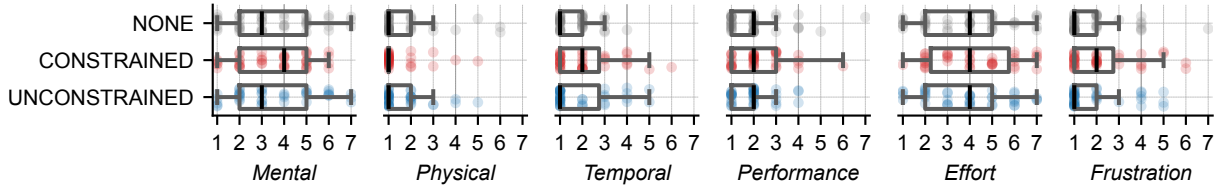


Figure 4.13: Questionnaire scores by condition after the reading portion of the experiment. Lower scores are better.

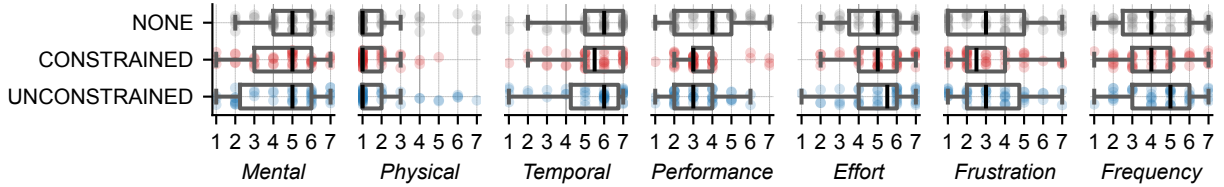


Figure 4.14: Questionnaire scores by condition after the test portion of the experiment. Lower scores are better for the TLX metrics (*Frequency* is simply descriptive, where higher scores indicate referring back to the story more).

4.4.3.1 Document Types

Seven participants (17%) in the CONSTRAINED condition said that highlighting was not necessary for a short story. One participant even noted feeling so engaged with the text that they forgot to highlight: “it helped me remember some important events, however, I now noticed that [when] reading the last parts of the study, I was so engrossed in it, I didn’t highlight that much” (P7). Two participants noted that their highlighting would be different for other types of documents, for example: “I was not concerned about the limit. It was not the type of fact-rich text that typically would be highlighted. For example, historical or medical type texts are ones that I would expect to highlight” (P6).

4.4.4 Summary

To summarize, our results suggest that constraining highlights can improve reading comprehension scores. Participants noted that having a word limit encouraged them to highlight only the most important points, and their highlights were in fact shorter and focused on highlighting key words like nouns, which is recommended by some university learning centres (e.g., [40]). This change in strategy did not increase reading time, nor did it increase mental demand, effort, or frustration when compared to an unconstrained ability to highlight.

4.5 DISCOURAGING EXCESSIVE HIGHLIGHTING

Inspired by these findings, we were curious to learn whether readers could be encouraged to highlight less through interaction techniques that make the act of highlighting more difficult with ‘soft’ constraints. As an initial exploration, we propose three new interaction techniques that purposely introduce friction

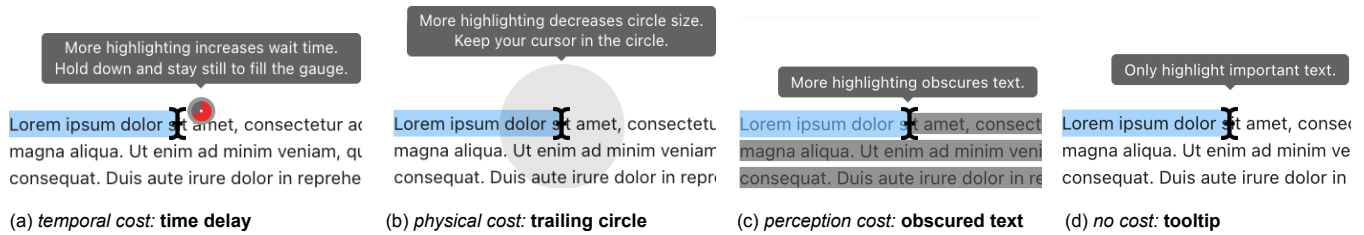


Figure 4.15: Interaction techniques that make the act of highlighting more (a) temporally, (b) physically, and (c) cognitively more demanding. We compare each to (d) a baseline, consisting of a tooltip.

to make highlighting more temporally, physically, and perceptually difficult to discourage excessive highlighting.

4.5.1 Time Delay

To increase the temporal cost of highlighting, this technique requires the user to press and hold the mouse button for a certain time to create the highlight. The time increases as more words are highlighted (Figure 4.15a). It works like this. First, as the user selects text they wish to highlight, instructions and a small gauge, visualized as a circular progress indicator, appear above the cursor. Second, after they stop moving the cursor, they must keep pressing the mouse button to fill the gauge. The fill is a red portion of the circle, and it increases every 100 ms. If the user moves the cursor as it is filled, the gauge is reset to zero. The gauge visualizes a *wait time*, after which the gauge turns green and the instructions tell the user to release the mouse button. After releasing, the highlight is created. If the user does not wait long enough before releasing, the highlight is not created.

The *wait time* is calculated by first summing the word count of text that has already been highlighted (*highlighted word count*) and the word count of text the user wishes to highlight (*new word count*). This sum, *total word count*, is linearly mapped to *wait time*, using a system-defined *target word count*, a *minimum wait time*, and a *maximum wait time*.

For example, suppose the user has already highlighted 10 words, the text selection is 5 words, the *target word count* is 30 words, and the wait times range from 0.1 s to 10 s. The *total word count* would be 15, representing 50% of the *target word count*, so the *wait time* would be approximately 5 s. Of course, the user may highlight more than the *target word count*, in which case the *wait time* could exceed the *maximum wait time*. This could be capped at the *maximum wait time*, or it could continue to grow until some *impossible time*, where the task becomes too frustrating or difficult. In our experiments, we explore values for these parameters.

4.5.2 Trailing Circle

This technique was designed to increase the physical cost of highlighting. The user must keep the cursor inside a circle that ‘trails behind’ and shrinks as the user highlights more words (Figure 4.15b). It works like this. First, when the

user moves the cursor to select text, a grey transparent circle and instructions appear to follow the user's cursor. As more words are selected, the circle dynamically shrinks to a new *circle diameter*. Conversely, if the user deselects text, the circle expands. The user must keep the cursor inside the circle at all times while selecting text. If they do this, the highlight is created when they release the mouse button. However, if the cursor leaves the circle, which can frequently happen with a small *circle diameter*, the circle turns red and 'pops' (visualized by expanding 300% and becoming transparent). If popped, the highlight is not created.

The trailing effect is created by filtering the circle position and diameter changes using the 1ϵ filter [35]. This ensures that the circle 'trails' behind the cursor and that large jumps in *circle diameter* do not cause the selection to fail. Much like the time delay technique, the *total word count* is linearly mapped to the *circle diameter* using the *target word count*, a *maximum circle diameter*, and a *minimum circle diameter*.

Building off the example above, if the *maximum circle diameter* is 100 px and the *minimum circle diameter* is 10 px, a *total word count* would be mapped to a *circle diameter* of 45 px. The *circle diameter* could be smaller than the *minimum circle diameter* if the *total word count* exceeded the *target word count*, and could become too challenging, for example, if the circle becomes smaller than the cursor itself. Therefore, it may be necessary to prevent the *circle diameter* from becoming smaller than an *impossible diameter*, values we explore in our experiments.

4.5.3 Obscured Text

This technique adds a perception cost, where the document text becomes more difficult to read, due to lower contrast of text and background colour, as the user highlights more words (Figure 4.15c). The technique works like this. First, when the user moves the cursor to highlight more words, the background for all text in the document is gradually darkened by an *obscure percent*. Only making the text background darker does not make selected text harder to read, so it is also lightened accordingly. When the user releases the mouse button, the document text is reset to the default white background and black text.

The *obscure percent* is determined through a linear mapping using the *total word count*, *target word count*, a *minimum obscure percent*, and a *maximum obscure percent*.

For example, if the *minimum obscure percent* is 5% and the *maximum obscure percent* is 85%, a *total word count* of 15 words would map to a 40% *obscure percent*. If the user highlights more than the *target word count*, the *obscure percent* could reach an *impossible percent* where the text background is fully black, or making the text completely illegible. We explore possible values for this in our experiments.

4.5.4 *Tooltip*

In addition to the aforementioned techniques that introduce friction, we also wondered whether a simple reminder could discourage excessive highlighting (Figure 4.15d). When the user selects text, a tooltip containing a suggestion to only highlight important text follows the cursor. Releasing the mouse button creates the highlight. This technique acts as a baseline.

4.5.5 *Pilot to Identify Design Parameters*

To determine optimal parameters for the minimum and maximum *wait time*, *circle diameter*, and *obscure percent*, we conducted an in-person within-subjects pilot experiment. Six participants from our institution (3 men, 3 women; ages 18-34) interacted with a custom interface that allowed them to try different settings for each technique by manipulating a slider. The interquartile ranges from the pilot results suggested that the following minimum and maximum values would be appropriate as they would be noticeable changes, but not too frustrating. For *wait time*, 0.4 s to 6.38 s; for *circle diameter*, 375 px to 63.75 px; and for *obscure percent*, 12.5% to 60%.

4.5.6 *Initial Evaluation*

Using these parameters, we first conducted a between-subjects experiment ($n=161$) on Prolific using all four interaction techniques. Much like Section 4.3, participants read a short story and highlighted text using a *TECHNIQUE*, with levels *TOOLTIP* ($n=40$), *TIME* ($n=41$), *TRAILING* ($n=40$), and *OBSCURE* ($n=40$). Then, they answered questions about the document in an open-book reading comprehension test. However, we made some changes to the protocol. We were primarily concerned with understanding the effect of *TECHNIQUE* on *Total Words Highlighted*, so we opted to use shorter, 550-word documents. Therefore, the mapping used 55 words as the *target word count*, representing roughly 10% of the 550-word document. We administered the reading comprehension task to participants immediately after they read the document. This was a 5-question multiple-choice reading comprehension test that they had to complete within 90 s. Both the stories and the tests were generated using GPT-4 and manually tweaked to ensure they were reasonably difficult. Our results suggested a significant effect of *TECHNIQUE* on *Total Words Highlighted* ($H = 8.87$, $p < .05$, $\eta_H^2 = .04$). However, post hoc tests were not significant ($.06 \leq p \leq .8$).

At this point, I had not discovered the passages created by Wallace et al. [175], which are shorter. We use these in Chapter 6.

4.5.6.1 *Refinement*

Based on these results, we refined our experimental design and the techniques. First, the experiment may have lacked statistical power due to a between-subjects design with four conditions. As our main focus was comparing each new *TECHNIQUE* to *TOOLTIP*, we decided to conduct multiple, within-subjects experiments.

Second, each technique could have been more difficult. In the initial evaluation, the ‘impossible’ values were the same as the ‘maximum’ values, but

Table 4.3: Participant demographics for Experiment 2.

Gender		Age		Education		English Language Proficiency	
Men	15	18-24	2	High School	3	Full Professional	1
Women	28	25-34	11	Some University (no credit)	14	Native or Bilingual	43
Non-binary	1	35-44	13	Technical Degree	6		
		45-54	7	Bachelor’s Degree	17		
		55-64	7	Master’s Degree	3		
		65-74	4	Beyond Bachelor’s (e.g., MD, JD)	1		

Document Reader Frequency		Highlight Frequency		Highlight Usage	
Daily	13	Daily	4	Remember Concepts	32
Weekly	9	Weekly	9	When Commenting	8
Monthly	11	Monthly	6	Concentration	17
Less than Monthly	9	Less than Monthly	16	Other	2
Never	2	Never	9		

these may have been too easy. For example, the *circle diameter* did not become smaller than 63.75 px, which is still reasonably large. To make the techniques harder, we changed them to use the minimum and maximum values from the pilot, and made the ‘impossible’ values more difficult. For *TIME*, the delay now ranged from 0.4 s to 7 s and extrapolated to a maximum of 30 s; for *TRAILING*, the circle diameter now ranged from 700 px to 30 px and extrapolated to 15 px, at which point keeping the cursor inside the circle becomes nearly impossible; and for *OBSCURE*, the level of obfuscation now ranged from 8% to 80% and extrapolated to 90%, at which point the text becomes nearly impossible to read.

Finally, the initial evaluation focused on assessing the number of highlighted words, but some techniques could potentially be too distracting and hinder reading comprehension. Our generated stories and reading comprehension tests were not validated for consistency and may have differed in difficulty, so we reverted back to the longer stories from easyCBM used in Experiment 1 (and set the *target word count* to 150 accordingly), and returned to a 24-hour gap between reading and completing the tests.

4.6 EXPERIMENT 2: INTERACTION TECHNIQUES

Learning from the previous between-subjects experiment, we conducted three related exploratory within-subjects experiments on Prolific.

4.6.1 Participants

Across all three experiments, we recruited 49 participants, but dropped 5 (10%) who experienced technical difficulties and those who did not attempt to highlight or answer any reading comprehension questions, leaving 44 valid responses (Table 4.3). Participants received \$15 in total and the top 25% received an additional \$3 bonus.

4.6.2 Task, Apparatus, and Design

The task was similar to Experiment 1, but this time, each participant read two short stories. They were randomly assigned to highlight text using TIME (n=19), TRAILING (n=9), or OBSCURE (n=16) for one story, and TOOLTIP for another. The order in which they experienced them was randomized. The reading interface was nearly identical to that of Experiment 1, but this time they could only highlight using the “Highlighter” tool. After reading and highlighting one document, they answered the same questions about their experience using the TECHNIQUE (all 0-100 scale). After reading and highlighting both documents, participants were shown their previously-entered responses, allowing them to refine them and better compare the two techniques [149]. Twenty four hours later, participants could complete both tests, which were presented in the same order they read them.

Studies that were conducted earlier use 1-7 scales, but newer studies use 0-100 scales to provide more granularity.

In Chapter 6, we randomize the test order too.

4.6.3 Results

The workload-related metrics suggest that the time delay, trailing circle, and obscured text made highlighting more difficult than seeing a tooltip, which was our objective. For OBSCURE, *Mental Demand* ($p = .009$) increased, all but *Temporal Demand* increased for TIME ($.0006 \leq p \leq .05$ for significant metrics), and all metrics increased for TRAILING ($.004 \leq p \leq .05$). However, even though the techniques were more difficult, we did not observe any meaningful trends. No TECHNIQUE was associated with increased *Reading Comprehension* when compared to TOOLTIP ($.48 \leq p \leq .92$). For *Total Words Highlighted*, there was a significant difference between TOOLTIP (MDN = 189, IQR = 264) and TRAILING (MDN = 113, IQR = 60; $W = 1$, $p = .007$, RBC = .96), however, it is likely that keeping the cursor inside the circle became almost impossible for participants to do, making it more of a ‘hard’ constraint than ‘soft.’

The sample size is admittedly small, so the study may lack power. However, because we did not observe any trends, we did not recruit more participants.

4.6.4 Summary

Overall, even though the techniques generally made highlighting more temporally, physically, and perceptually more demanding, we did not observe differences in the number of highlighted words for the time delay and obscured text techniques. For the trailing circle technique, we did observe a reduction in the number of highlighted words. None of the techniques were associated with increased reading comprehension.

4.7 DISCUSSION

We showed that imposing a word limit on the number of highlightable words can improve reading comprehension. However, when we tried to use ‘soft’ constraints that introduce friction into the act of highlighting, we did not observe any improvements in reading comprehension. We discuss research directions based on our work for the broader HCI community and the limitations of our work.

4.7.1 *Research Directions for HCI*

A text highlight constraint is a very simple concept that could be integrated into existing document readers like Adobe Acrobat and macOS Preview. However, there remains open questions and design decisions for HCI researchers.

4.7.1.1 *Identifying Constraints for Different Documents and Task Environments*

During our pilot studies for Experiment 1, we identified 150 words (roughly 10% of the document word count) as a promising constraint for the types of short stories and the type of task we selected: a reading comprehension test with additional pressure from a short time limit. However, as suggested by participants and our own pilot testing (see [Section 4.3.3](#) for details), this is not a universal solution as different documents and tasks will require different levels of constraint. One possibility is to allow users to set their own levels of constraint for individual documents, but a more ambitious goal would be to analyze document characteristics to calculate an optimal level of highlight constraint, possibly using the capabilities of LLMs. Constraints could be imposed relative to the text structure, for example, allowing only 5 highlights per section, or allowing more highlights for certain types of sections, like the results section of an academic paper. This might encourage readers to highlight more consistently throughout a document, which was something our participants seemed to struggle with.

Studying for a test has a clear objective, but text may be highlighted for a variety of reasons, many of which are more exploratory or ambiguous [39]. For example, knowledge workers may frequently switch between broadly capturing text to gather information and narrowing or filtering text to create meaningful insights [79, 142]. Similarly, when collaborating with others, a reader will likely highlight more text initially before filtering their personal highlights to share with others [116]. A text highlight constraint should adapt to suit the nature of these tasks, allowing for more words to be highlighted for exploratory or ambiguous tasks, and less words for narrowing or filtering tasks.

Text highlight constraints could help readers learn. For example, in a classroom setting, teachers could set levels of constraints to encourage better study habits among students. The level of constraint could even act as training, where a document reader first learns current highlighting behaviours, then gradually imposes a more constrained word limit over time to help readers to develop better highlighting strategies.

4.7.1.2 *Integration with Existing Features*

Recall 37 participants (29%) reported previous highlighting experience linked to adding a text comment ([Table 4.1](#)). Imposing word limits on other types of document annotations like comments may have similar benefits, but it is unclear how text highlight constraints should be combined with text comment constraints since they are often simultaneous. One option is to have separate word limits for highlights and comments, allowing for a separation of concerns. Another option is to have a combined word limit. Both options

increase user effort: either keeping track of multiple limits or allocating words across these two features.

Chapter 6 extends commenting capabilities, however, it does not impose word limits.

4.7.1.3 Interaction Techniques

Our assumption when designing and evaluating new interaction techniques was that making the act of highlighting more difficult would disrupt fast and automatic System 1 thinking for slower and deliberate System 2 thinking ([dual process theory](#); [Section 2.3](#)). This would encourage participants to self-reflect on their highlighting and highlight fewer words. However, this did not seem to happen: participants generally highlighted the same number of words as just seeing a tooltip. When they did highlight fewer words for the trailing circle technique, this did not translate to better reading comprehension scores.

There are several factors that may explain these results. One is that just seeing a tooltip is enough to [nudge](#) highlighting behaviour. To verify this, we compared TOOLTIP to UNCONSTRAINED and CONSTRAINED using Mann-Whitney U tests. We did not observe any significant differences between TOOLTIP and UNCONSTRAINED for *Total Words Highlighted* ($p = .62$) and *Reading Comprehension* ($p = .82$), but we did observe significant differences between TOOLTIP and CONSTRAINED for both *Total Words Highlighted* ($p < .001$) and *Reading Comprehension* ($p = .04$). This suggests that a hard word limit is associated with fewer words highlighted and better reading comprehension scores, and that just seeing a tooltip is not enough to discourage excessive highlighting.

Another possibility is that knowing there was a hard limit on highlightable words triggered a deeper engagement with the text that was more important for reading comprehension than just reducing the number of highlighted words. As another check, we grouped participants in Experiment 1 by how many words they actually highlighted, regardless of condition: 0 words (NONE; $n=44$), less than or equal to 150 words (150; $n=56$), and more than 150 words (MORE; $n=27$). Participants did significantly better with 150 (MDN = 14, IQR = 5) than NONE (MDN = 10, IQR = 7; $p = .04$), but there were no significant differences between 150 and MORE (MDN = 14, IQR = 6). This suggests that other, more psychological factors, such as the ability to recognize keywords and phrases, may be more important to increasing reading comprehension than just word count, and that this was encouraged through hard constraints but not soft constraints.

Future work could explore other types of interaction techniques or further tune the ones we proposed. For example, our implementation linearly mapped word count to *wait time*, *circle diameter*, and *obscure percent*, but non-linear mappings may be more appropriate.

4.7.2 Limitations

We tested multiple stories for improved external validity, but this reduced the number of participants per story for each condition. This made it difficult to analyze the raw highlight text and differences between conditions in greater depth. Repeating Experiment 1 with a single story or greatly increasing the number of participants would allow us to learn what text is commonly highlighted between participants and give better insights into

how people change strategies when highlighting under a word limit. Our results suggest that constrained highlighting could teach readers effective highlighting techniques, but this could be further validated by comparing it to an unconstrained highlighting condition in which participants learn effective highlighting techniques before reading a document.

Although we tried our best to mimic what it would be like to study for a test, our experimental setup may be lacking in ecological validity. Specifically, Lonka et al. [111] note that study strategies used during an experiment may be different than those used when studying for an actual exam. The documents we selected allowed for high internal validity and roughly correspond to something a student may face in an English course. However, there are other types of documents where text-marking is arguably even more useful, possibly with additional text-marking tendencies (e.g., non-fiction articles). A longitudinal study of text highlight constraints within a real educational setting would further validate and extend our findings.

Prior work suggests that people with lower reading abilities struggle to identify key concepts to highlight [10]. Although we conducted Experiment 1 with a large population with a diverse educational background, we did not formally measure reading ability. This would have required additional time-consuming tasks, like the Nelson-Denny reading test, which would greatly increase the duration and fatigue of our experiment. It is likely that constrained highlighting would be more difficult and mentally demanding for this population, but perhaps once mastered, they may experience the greatest improvements in reading comprehension.

4.8 CONCLUSION

Using a large-scale, between-subjects experiment, we show that a text highlight constraint can improve reading comprehension scores when compared to not highlighting anything or unlimited highlighting. Our work validates theories in psychology, which state that being more selective when highlighting text improves recollection. At its core, the idea of constraining text highlights is incredibly simple. However, we believe that incorporating it into existing document reader software is an ‘easy win’ that can help people become better learners by forcing them to be more selective and intentional with their highlights without the need for lengthy and time-consuming self-regulation training. We then designed and implemented interaction techniques that make highlighting temporally, physically, and perceptually more difficult, yet these techniques did not improve reading comprehension, suggesting that ‘hard’ constraints may be more effective. We hope our work will inspire new features and interactions within document reader software that are designed around constraints.

REQUIRING LONGER PROMPTS TO INCREASE PSYCHOLOGICAL OWNERSHIP WHEN WRITING WITH AI

In the previous chapter, we limited how many words readers could highlight in document reader software and found that this increased reading comprehension scores. This constraint restricted user actions *after a certain point*, specifically, *when* they reached 150 highlighted words. In this chapter, we continue to examine the effects of constraining *when* interaction can occur, by restricting interaction *before a certain point*. Specifically, by requiring users to write longer prompts to increase [psychological ownership](#).

5.1 MOTIVATION

People often write collaboratively, which can lead to many benefits, like improved text quality [133, 189]. However, it can also lead to negative outcomes, such as a loss of [psychological ownership](#) [24, 25, 36]. This is a concept about feelings of the text belonging to the writer, regardless of legal ownership [139]. Preserving feelings of psychological ownership is important when writing collaboratively, as prior work suggests that losing too much psychological ownership can deter people from writing together in the future [36].

People are increasingly writing ‘collaboratively’ with generative AI services like ChatGPT. For example, author Rie Kudan, winner of the 170th Akutagawa Prize in Japan, used ChatGPT to write her award-winning book [42]. Given this new form of collaborative writing, understanding how psychological ownership is affected by generative AI writing assistants is crucial. It seems intuitive that writing with a generative AI assistant will result in lower feelings of psychological ownership than writing something independently, which has been corroborated by early work investigating the roles and designs of AI writing assistants (e.g., [54, 104, 108]). However, we argue that more nuance of understanding is needed. Unlike collaboratively writing with another person, whose actions cannot be controlled, someone collaborating with a generative AI assistant can exert more control over the output through their prompts, which can vary in *length*. Someone may write a short prompt that lacks detail about the content or style, or may write a long prompt to include these details ([Figure 5.1](#)). Writing longer and detailed prompts should result in generated text that is more representative of these additional details, and including these details requires an investment of time, energy, and sense of self; all of which are important for psychological ownership [138, 139, 178]. Therefore, someone who invests more by writing longer prompts may have increased feelings of psychological ownership than someone who writes less, and it is possible that writing longer prompts may even lead to similar levels of psychological ownership as independent writing. Despite the potential influence of prompt length on psychological ownership, this has not been empirically investigated.

We conducted two experiments to answer the following research question: *how much psychological ownership do people feel over writing produced by generative*

	Where	When
R	Ch. 3	Ch. 4
W		Ch. 5
R+W	Ch. 6	

*This project explores constraining **when** prompts can be submitted while **writing** a document with AI assistance.*

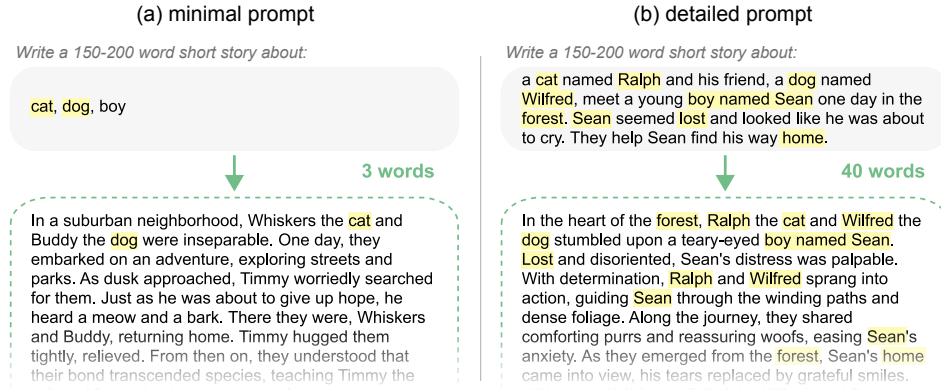


Figure 5.1: Writing a short story with (a) a minimal prompt with only 3 words and (b) a detailed prompt with 40 words. Yellow highlight shows similarities between the original prompt and the resulting story.

AI, and how does prompt length affect it? During the first experiment, participants wrote 150-200-word short stories either completely by themselves or by writing prompts of varying lengths enforced through word minimums and maximums, effectively constraining the prompt length. Results showed that as participants wrote more to meet higher word minimums, they had higher levels of psychological ownership. Specifically, participants included more details about the story plot in their prompts, which resulted in more similarities between their prompts and their stories, and required more thought. The second experiment used even higher word minimums, yet psychological ownership plateaued.

These experiments required users to write longer prompts using word minimums. Though valuable for increased experimental control, this approach may not work well in practice, where users desire some flexibility. As such, we examine an important follow-up research question: *how can we encourage users to write longer prompts?* We propose two simple changes to the prompt entry interface that introduce *time* and *motor* costs based on prompt length. Specifically, the user must press and hold the prompt submission button for some time or repeatedly move a slider up and down before their prompt is submitted. The time spent waiting and the number of slider movements are both functions of prompt length, meaning if someone writes a longer prompt, they experience less delay or do not have to move the slider as much.

Using a similar creative writing task, we first explored the effectiveness of a time cost and found that 20- and 60-second delays were associated with longer prompt lengths than having no delays. Learning from these results and the experimental protocol, we then compared the time cost to the motor cost. When compared to a baseline interface with no augmentations, participants felt more mental demand, physical demand, effort, and frustration, but wrote significantly longer prompts and felt more psychological ownership toward the resulting stories. To encourage participants to generate more ideas [160] and write even longer prompts, we further augmented these techniques to display AI-generated suggestions every 8 seconds, triggered after one second of inactivity. A follow up study showed that this significantly increased

prompt length for the augmented prompt entry conditions, but did not further improve psychological ownership.

The effects of writing with a generative AI assistant on psychological ownership is important for the HCI community to understand, as the community is developing writing systems with generative AI support that offload work from the user. While this may be beneficial for productivity, users may come to feel negatively about such systems and stop using them over time, especially for tasks where the end product should be ‘theirs.’ Our work contributes empirical and quantitative findings on the effects of writing with a generative AI assistant on psychological ownership and how this is impacted by prompt length, and two simple yet effective approaches for increasing prompt length through time and motor costs that can be integrated into current chat-based AI interfaces.

5.2 BACKGROUND AND RELATED WORK

It has been long understood that people feel possessive towards objects, ideas, places, and creations. In psychology, these possessive feelings are referred to as *psychological ownership* [138, 139]. Many factors contribute towards increased feelings of psychological ownership, like the investment of time, energy, and sense of self, which may explain why people feel possessive towards creative work [138, 139, 178]. Feeling psychological ownership can have positive outcomes, like an increased sense of responsibility and citizenship behaviour [139], but can negatively impact collaborative tasks [138]. People often write collaboratively, which can improve text quality [133, 189], however, writers often feel psychological ownership over jointly-written text [25, 36, 71, 171]. Although some types of collaborative writing can improve psychological ownership, such as synchronous collaborations with more communication [17], fears of losing psychological ownership can often lead to territorial behaviours [101] or avoiding future collaboration [36].

Psychological ownership has been identified as important to the design [102] and evaluation [158] of AI writing systems. Biermann et al. [15] conducted interviews with hobbyist and professional writers to better understand how they write and the challenges they face while writing. Participants commented on mock-up user interfaces designed for different AI assistant roles, such as using the AI to receive suggestions for new text or to generate dialog for different characters. Psychological ownership was an important factor for these writers, and some designs were poorly received because they “overstep the boundaries” of the human writer. Draxler et al. [54] asked participants to write postcards entirely by themselves or to edit or choose postcards pre-written by AI before ‘signing’ the postcard. Participants did not feel psychological ownership towards text that was pre-written by AI, however, they still did not list “AI” as an author, suggesting the presence of an *AI Ghostwriter Effect*. Li et al. [108] asked participants to write persuasive essays and stories with an AI assistant, where the human and the AI could be the primary writer or the editor. Having the AI as the primary writer lowered psychological ownership, however, participants still valued AI assistance as they were willing to forgo some payment to receive help. These early evaluations suggest that psychological ownership is lowered when collaborating

with AI, however, they focused on the role of the AI and did not explore other factors that contribute towards lower psychological ownership in depth.

There are many writing systems that leverage AI and LLMs (e.g., [78, 191]), but we focus on those that evaluated psychological ownership. Dramatron [120] helps writers create theatre screenplays and scripts, but writers reported low feelings of psychological ownership over the resulting pieces. GhostWriter [186] allows writers to refine writing style preferences through highlights, but psychological ownership was the most varied and lowest scoring metric. Metaphoria [62] suggests metaphors while writing poems, but writers felt less psychological ownership when using it, especially when “the suggestions were particularly good.” The type of suggestion may also play a role. For example, Dhillon et al. [51] found that psychological ownership decreased when writers received entire paragraphs or sentences as suggestions from an AI writing assistant. Lee et al. [103] suggested that as writers receive more suggestions from AI, they write less and feel less psychological ownership. However, they note that these results were preliminary and more research is needed to further validate these findings.

It has been hypothesized that increased control is desirable for human-AI collaborations (e.g., [131, 148, 152]), and that having more control can increase feelings of psychological ownership while writing. For example, Lehmann et al. [104] compared AI-generated suggestions that the user had to manually accept to suggestions that were automatically inserted into the text body. They found that manually-accepted suggestions was associated with higher feelings of psychological ownership, which may be due to the increased control at the composition stage.

To summarize, prior work has shown that psychological ownership is important to writers who write collaboratively with humans and with AI. Although some AI writing systems have included features to give writers more control to increase psychological ownership, the effects of more fundamental properties, such as the prompt length, have yet to be explored.

5.3 EXPERIMENT 1: PROMPT LENGTH

The goal of this experiment is to understand the impact of writing with generative AI on psychological ownership and how prompt length affects it. Through a within-subjects experimental design, participants wrote short stories with generative AI assistance by writing prompts of varying word lengths, which were enforced through word minimums and maximums for increased experimental control: 3 words, 50-100 words, and 150-200 words. As a baseline, participants also wrote a 150-200 word short story without any AI assistance.

5.3.1 *Participants*

We recruited 41 participants through the crowdsourcing experiment service, Prolific. Participants were restricted to the United States and Canada, those who had completed at least 2,500 tasks, and those who had an approval rating greater than 98%. We manually examined all open-ended responses and user

Table 5.1: Demographic information for Experiment 1.

Gender		Age		Education		Creative Writing Frequency		Prompt Engineering Familiarity	
Men	14	25-34	10	Some University (no credit)	4	Daily	1	Extremely	3
Women	16	35-44	7	Technical Degree	1	Weekly	7	Moderately	3
Non-binary	1	45-54	6	Bachelor’s Degree	21	Monthly	7	Somewhat	3
		55-64	6	Master’s Degree	5	Less than Monthly	11	Slightly	7
		65-74	2			Never	5	Not at All	15

ChatGPT Frequency		Weekly ChatGPT Use		Other AI Services				AI Usage			
Daily	3	0 times	7	Nothing	3	Gemini	9	Nothing	1	Generate Images	12
Weekly	14	1-5 times	13	GPT-2	6	Grammarly	9	Write Emails	11	Get a Definition	9
Monthly	4	5-10 times	8	GPT-3	12	Stable Diffusion	2	Write Papers	2	Explore a Topic	17
< Monthly	9	10-30 times	2	GPT-4	6	Dall-E	4	Write Stories	6	Brainstorming	25
Never	1	30+ times	1	Copilot	2	Midjourney	5	Edit Text	14	Find References	8
				Bing	12	Other	6	Generate Code	3	Clarification	7
								Edit Code	3	Translate Text	5
								Debug Code	3	Other	4

input to identify fraudulent responses [153] and filtered out participants who experienced technical issues with our interface and those who cheated by padding out their prompts or stories with dummy text to meet the word minimums and those who did not make enough keystrokes to generate their user input, suggesting they copied and pasted content from another source. In total, 10 participants (24%) were excluded, leaving 31 valid responses (Table 5.1). All self-reported being proficient at reading and writing in English (all ≥ 5 on a 1-7 scale). Participants received \$15 as remuneration.

5.3.2 Apparatus and Task

The experiment software was a Node.js and React web application (Figure 5.2). The main writing interface had a toolbar on the top and a side-by-side view underneath. The top toolbar contained instructions for the trial and three nouns that the story had to be about (i.e., a unique and randomly assigned “triad” from a set of nine triads adapted from Foley et al.’s text composition task [59]), which is useful for encouraging creative thinking on the spot during experiments [55]. Writing independent short stories is an ideal task for this type of experiment as creative writing requires a high personal investment and expression of self, which are important for psychological ownership, yet it provides increased control for an experiment.

A text box on the left provided a space for participants to write a prompt for the GPT-3.5 Turbo model. In a baseline condition, they wrote a complete story. Copying and pasting within the text box were disabled to prevent cheating, but this may not work across all browsers, so keystroke-level activity was also logged to verify response validity. A word counter above the text box displayed the number of words in the text box. Once the word count satisfied the requirements for the condition (i.e., the word count was between the minimum and maximum), a “Finalize” button could be pressed, causing the story

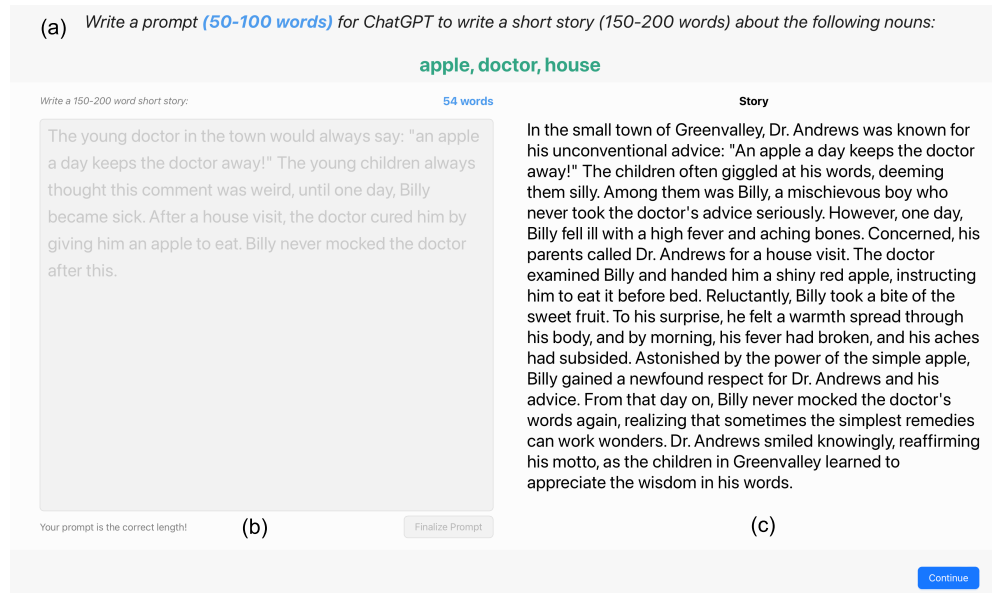


Figure 5.2: Experimental interface (shown with AI assistance): (a) top toolbar showing the instructions and three nouns the story had to be about; (b) a text box where the participant wrote their prompt or short story; (c) the story written by themselves or with AI assistance.

to 'type out' on the right side of the screen, following the behaviour of LLM AI systems like ChatGPT. The "Finalize" button could only be pressed once. The story was either generated by the model by including the participant's prompt within an engineered prompt, or it was the same text written by the participant in the baseline. The same typing effect was used in both cases, and this provided an opportunity for participants to read the story. After the story was fully typed, a "Continue" button could be pressed to end the trial.

5.3.3 Procedure

Participants received a link to the web application through Prolific and the experiment was restricted to laptop and desktop devices. After entering basic demographic information (questions were adapted from Masson et al. [118]) and reading instructions, they were shown the writing interface and were asked to write a short story or a prompt about the three presented nouns. After the writing stage, they answered questions about their story and the writing experience. They repeated this for all conditions before answering questions about their overall thoughts and preferences. The entire experiment took approximately one hour.

5.3.4 Choice of Prompt Lengths

We were interested in examining the effects of prompt length, believing that being required to write longer prompts will increase the prompt level of detail and improve psychological ownership. As such, we needed to test a range

of prompt lengths representing key cases, like a story generated with very little guidance from the writer, a story heavily guided by the writer, and a story written solely by the writer. At the same time, we had to ensure that the experiment duration was not too long for participants. Therefore, our choice of word minimums (and word maximums) was guided by the duration of conditions requiring more writing from the participant (e.g., writing a story without any AI assistance). We ultimately decided that the final stories for all conditions had to be 150-200 words, so that the experiment was not too time-consuming but the stories were still long enough to be fully developed. With a 150-200 word final story in mind, we chose conditions to examine the effects of prompt length and a baseline:

- **3 words with GPT**, representing 1% of the maximum final story word count. This best approximates the case where the story is fully generated by an AI assistant.
- **50-100 words with GPT**, representing 25-50% of the maximum final story word count. The story is generated by an AI assistant but has some guidance from the participant.
- **150-200 words with GPT**, representing 75-100% of the maximum final story word count. The story is generated by an AI assistant, but with a lot of guidance from the participant.
- **150-200 words without GPT**, a baseline representing the case where the story is fully written by the participant.

5.3.5 Design

This is a within-subjects design with one primary independent variable, *CONDITION* (levels: AI-3, AI-50, AI-150, NO-AI; the number representing the word minimum). *CONDITION* was randomly assigned. For each *CONDITION*, the primary measures were 10 subjective question scores, all interval numeric scores within a 1-7 range. Four questions represented metrics related to psychological ownership and the phrasing was similar to prior work [36, 124]: *Personal Ownership*, *Responsibility*, *Personal Connection*, *Emotional Connection*. We average these four scores to create a composite measure, *Psychological Ownership*, which is a common technique in psychology papers. The internal consistency reliability score, which describes how consistently different items on a questionnaire describe the same underlying concept, was calculated using Cronbach’s alpha, and the reliability was very high ($\alpha = 0.96$), indicating that *Psychological Ownership* was a reliable composite measure for data analysis. Six questions represented metrics from the NASA-TLX: *Mental Demand*, *Physical Demand*, *Temporal Demand*, *Performance*, *Effort*, and *Frustration*. Objective metrics were calculated from logs: the time taken in minutes to write the prompt or story (*Duration*); the number of words in each (*Word Count*); and the semantic text similarity between the final story and the participant’s prompt (*Text Similarity*), which was calculated using Google’s Universal Sentence Encoder (0-1 range; 1 being identical texts) [37].

Table 5.2: Omnibus and post hoc statistical tests for Experiment 1: (a) *Psychological Ownership*, (b) *Duration*, (c) *Word Count*, (d) *Text Similarity*, (e) *Mental Demand*, and (f) *Effort*.

(a) <i>Psychological Ownership</i>					(b) <i>Duration</i>					(c) <i>Word Count</i>				
$\chi^2_{3,N=31} = 78.01, p < .001, W = .84$					$\chi^2_{3,N=31} = 35.86, p < .001, W = .38$					$\chi^2_{3,N=31} = 85.87, p < .001, W = .92$				
<i>comparisons</i>		<i>p-value</i>		<i>RBC</i>	<i>comparisons</i>		<i>p-value</i>		<i>RBC</i>	<i>comparisons</i>		<i>p-value</i>		<i>RBC</i>
AI-3	AI-50	< .001	***	.72	AI-3	AI-50	.87	<i>n.s.</i>	.22	AI-3	AI-50	< .001	***	1.0
AI-3	AI-150	< .001	***	.75	AI-3	AI-150	.02	*	.52	AI-3	AI-150	< .001	***	1.0
AI-3	NO-AI	< .001	***	.99	AI-3	NO-AI	.01	**	.58	AI-3	NO-AI	< .001	***	1.0
AI-50	AI-150	.01	**	.25	AI-50	AI-150	< .001	***	.67	AI-50	AI-150	< .001	***	1.0
AI-50	NO-AI	< .001	***	.83	AI-50	NO-AI	< .001	***	.76	AI-50	NO-AI	< .001	***	1.0
AI-150	NO-AI	< .001	***	.65	AI-150	NO-AI	.24	<i>n.s.</i>	.16	AI-150	NO-AI	.005	**	.4

(d) <i>Text Similarity</i>					(e) <i>Mental Demand</i>					(f) <i>Effort</i>				
$\chi^2_{3,N=31} = 99.00, p < .001, W = 1.00$					$\chi^2_{3,N=31} = 60.21, p < .001, W = .65$					$\chi^2_{3,N=31} = 49.99, p < .001, W = .54$				
<i>comparisons</i>		<i>p-value</i>		<i>RBC</i>	<i>comparisons</i>		<i>p-value</i>		<i>RBC</i>	<i>comparisons</i>		<i>p-value</i>		<i>RBC</i>
AI-3	AI-50	< .001	***	.7	AI-3	AI-50	< .001	***	.57	AI-3	AI-50	< .001	***	.76
AI-3	AI-150	< .001	***	.89	AI-3	AI-150	< .001	***	.82	AI-3	AI-150	< .001	***	.76
AI-3	NO-AI	< .001	***	1.0	AI-3	NO-AI	< .001	***	.86	AI-3	NO-AI	< .001	***	.86
AI-50	AI-150	< .001	***	.55	AI-50	AI-150	< .001	**	.37	AI-50	AI-150	.16	<i>n.s.</i>	.17
AI-50	NO-AI	< .001	***	1.0	AI-50	NO-AI	< .001	***	.44	AI-50	NO-AI	.02	*	.33
AI-150	NO-AI	< .001	***	1.0	AI-150	NO-AI	.33	<i>n.s.</i>	.04	AI-150	NO-AI	.13	<i>n.s.</i>	.12

5.4 EXPERIMENT 1: RESULTS

We conducted Shapiro-Wilk and Mauchly's tests to determine the normality and sphericity of the data, and all metrics significantly deviated from a normal distribution or did not have sphericity. Therefore, we use a Friedman test and Wilcoxon signed-rank tests, with Holm's corrections for multiple comparisons; and Spearman's correlations. Table 5.2 shows details of statistical tests.¹

5.4.1 *Psychological Ownership*

Overall, writing more words improved feelings of psychological ownership (Figure 5.3). A significant effect of CONDITION on *Psychological Ownership* and post hoc tests revealed that AI-3 was associated with the lowest scores (MDN = 1.25, IQR = 1.25), followed by AI-50 (MDN = 3.75, IQR = 2.12), AI-150 (MDN = 4.75, IQR = 1.75), and NO-AI (MDN = 6.5, IQR = 1).

¹ Kendall's *W* acts as the effect size for the Friedman test.

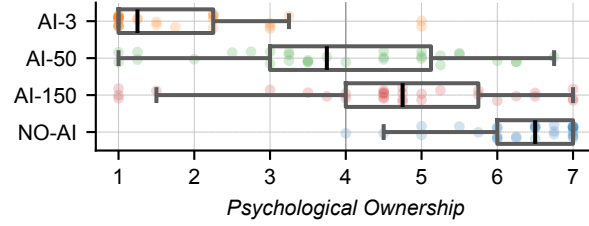


Figure 5.3: *Psychological Ownership* by *CONDITION*. Higher scores correspond to higher feelings of *Psychological Ownership*.

5.4.2 Writing Activities

To better understand factors that may have contributed to these results, we examine participants' writing activities and task workload. I independently grouped the open-ended responses as the data was straightforward given the number of participants and the short response length (2-3 sentences each).

5.4.2.1 Duration

We anticipated that spending more time writing may lead to higher *Psychological Ownership*, but they were only moderately positively correlated ($r_s = .45, p < .001$). A significant effect of *CONDITION* on *Duration* and post hoc tests revealed that AI-3 (MDN = 1.85, IQR = 5.87) and AI-50 (MDN = 3.25, IQR = 2.47) were faster than AI-150 (MDN = 6.99, IQR = 6.08) and NO-AI (MDN = 7.89, IQR = 4.46; Figure 5.4).

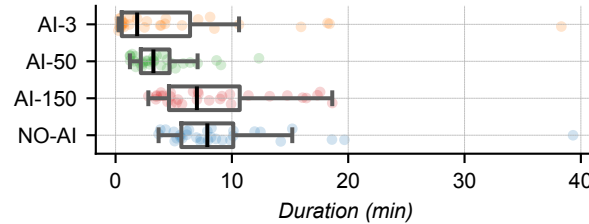


Figure 5.4: *Duration* by *CONDITION*.

5.4.2.2 Word Count

We observed a strong positive correlation between *Word Count* and *Psychological Ownership* ($r_s = .72, p < .001$), suggesting that as participants wrote more, they felt more psychological ownership over their writing. As expected, we observed a significant effect of *CONDITION* on *Word Count* with post hoc tests showing participants wrote fewer words with AI-3 and AI-50 (MDN = 60, IQR = 28.5) than they did with AI-150 (MDN = 157, IQR = 10) and NO-AI (MDN = 168, IQR = 28.5; Figure 5.5). Despite having the same word limits, we also observed a significant difference between AI-150 and NO-AI. One

possibility is that participants encountered a mental block with AI-150, as some mentioned how it was “*more taxing to write [a] longer prompt*” (P31) and how they “*couldn’t think of what [...] to write about*” (P21).

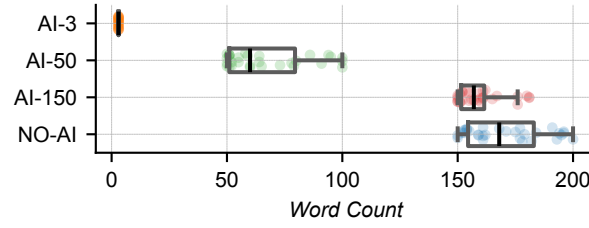


Figure 5.5: Word Count by CONDITION.

5.4.2.3 Text Similarity

Overall, writing longer prompts was associated with more semantic text similarity with the generated story. There was a strong positive correlation between Word Count and Text Similarity ($r_s = .82, p < .001$), and a significant effect of CONDITION on Text Similarity, revealing that prompts written with AI-3 were the least similar to the final story (MDN = .30, IQR = .12), followed by AI-50 (MDN = .41, IQR = .16), AI-150 (MDN = .61, IQR = .15), and NO-AI (Figure 5.6). Much like Word Count, having increased Text Similarity was strongly positively correlated with higher Psychological Ownership ($r_s = .75, p < .001$).

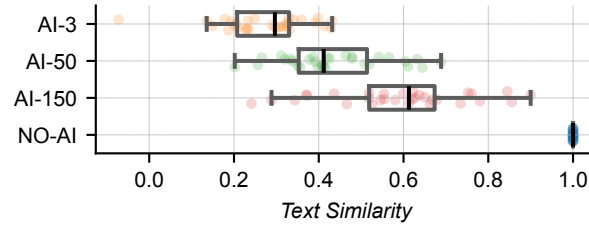


Figure 5.6: Text Similarity by CONDITION. Higher scores refer to more similarities between the input and output (1 = identical texts).

5.4.2.4 Prompt Strategies

We observed participants adopting several strategies with their prompts (Table 5.3). As expected, most participants wrote the three specified nouns for AI-3, so we only consider AI-50 and AI-150. Only 4 prompts (6%) purely consisted of instructions without any details about the story plot; 25 (40%) purely consisted of story plot details and resembled partially complete stories; and majority (33, 53%) were a mix of instructions and story details. As would be expected, it seems like requiring participants to write longer prompts encouraged greater levels of details about the story, something that 11 (35%) participants noted, for example: “*when I provide more words, I see that [it]*

Table 5.3: Example prompt types submitted by participants (examples from AI-50 condition).

Plot	Mix	Instructions
<i>"The girl asked her grandma for a dog every time she saw her. The grandma kept saying no, until finally she surprised her granddaughter with a dog for Christmas that year. The girl had tears of joy that she finally got a dog and was in disbelief that it was happening." (P18)</i>	<i>"Create an average word paragraph using the following nouns: House, grandpa and mug. Start with a setting of grandpa sitting in a rocking chair on the porch. Have grandpa tell a quick story to his 8 year old grand daughter named Lisa. End story with grandpa going into the house." (P15)</i>	<i>"Write a short story about these three nouns: Boy, boat, cat. The story should be 150 - 200 words in length. The whole story should be about these three nouns. Write a story that can be read to children, so the words should be as simple as possible so children could understand." (P26)</i>

produces a story that is almost the same as mine" (P1). Of these participants, 8 (25%) noted how this increased feelings of psychological ownership, for example: "with longer prompts, I could influence what was on display and felt more ownership [towards] the work" (P13).

5.4.2.5 Task Workload

The longest prompt and writing the full story were more mentally demanding and the shortest prompt required the least effort (Figure 5.7), but we did not observe meaningful trends for other task workload factors. For *Mental Demand*, a significant effect of CONDITION and post hoc tests revealed that AI-3 (MDN = 1, IQR = 0) and AI-50 (MDN = 3, IQR = 3) were less mentally demanding than AI-150 (MDN = 4, IQR = 4) and NO-AI (MDN = 5, IQR = 2.5; Figure 5.7a). There were no significant differences between AI-150 and NO-AI. *Mental Demand* was strongly positively correlated with *Psychological Ownership* ($r_s = .65, p < .001$).

There was a significant effect of CONDITION on *Effort*. Unsurprisingly, post hoc tests showed that AI-3 required the least *Effort* (MDN = 3, IQR = 0). Although AI-50 (MDN = 3, IQR = 2.5) required less *Effort* than NO-AI (MDN = 5, IQR = 2), there were no significant differences between AI-50 and AI-150 (MDN = 4, IQR = 3.5) or between AI-150 and NO-AI (Figure 5.7b). *Effort* was strongly positively correlated with *Psychological Ownership* ($r_s = .66, p < .001$). The relationship between *Mental Demand* and *Effort* and *Psychological Ownership* makes sense; it has been long understood that investing more energy

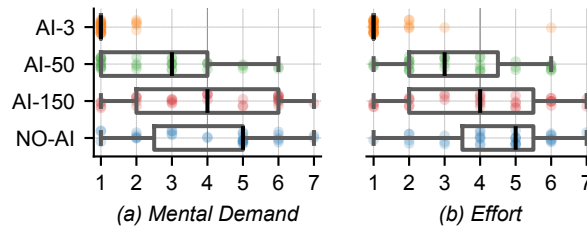


Figure 5.7: (a) *Mental Demand* and (b) *Effort* by CONDITION. Lower scores correspond to lower *Mental Demand* and *Effort*.

Table 5.4: Demographic information for Experiment 2.

Gender		Age		Education		Creative Writing Frequency		Prompt Engineering Familiarity	
Men	12	18-24	3	High School Diploma	3	Daily	2	Extremely	1
Women	19	25-34	11	Some University (no credit)	8	Weekly	3	Moderately	3
Non-binary	1	35-44	4	Technical Degree	2	Monthly	7	Somewhat	7
Other	1	45-54	11	Bachelor’s Degree	18	Less than Monthly	12	Slightly	12
Unknown	1	55-64	4	Master’s Degree	2	Never	10	Not at All	11
		Unknown	1	Doctorate Degree	1				

ChatGPT Frequency		Weekly ChatGPT Use		Other AI Services				AI Usage	
Daily	10	0 times	8	Nothing	6	Gemini	7	Nothing	2
Weekly	8	1-5 times	12	GPT-2	6	Grammarly	8	Generate Images	14
Monthly	6	5-10 times	4	GPT-3	12	Stable Diffusion	3	Write Emails	14
Less than Monthly	6	10-30 times	6	GPT-4	6	Dall-E	9	Write Papers	4
Never	4	30+ times	4	Copilot	0	Midjourney	6	Explore a Topic	22
				Bing	10	Other	4	Write Stories	5
								Brainstorming	19
								Edit Text	15
								Find References	5
								Generate Code	3
								Clarification	16
								Edit Code	3
								Translate Text	5
								Debug Code	3
								Other	3

and self improves psychological ownership [138, 139, 178], which requires more thought and effort.

5.4.3 Summary

To summarize, our results suggest that requiring people to write longer prompts increases psychological ownership for their writing. By writing more, prompts became more similar to the resulting stories, likely because writers were encouraged to include more details about the story and its plot, which required more mental demand. An interesting question is: does this effect plateau with even longer prompts, perhaps even longer than the generated text?

5.5 EXPERIMENT 2: PLATEAU

To better understand this, we conducted a follow-up experiment with 39 new participants. Five (13%) were removed from the analysis for cheating or because they experienced technical issues, leaving 34 participants (Table 5.4). All participants self-reported being proficient at reading and writing in English (all ≥ 5 on a 1-7 scale). The protocol and measures were the same as Experiment 1 ($\alpha = 0.92$ for *Psychological Ownership*), but participants had to write prompts that were 150-200, 175-200, and 200-250 words long. The experiment was slightly longer at roughly 90 minutes and participants received \$22.50 as remuneration.

5.5.1 Choice of Prompt Lengths

Two conditions were identical (AI-150 and NO-AI), and two new word minimums and maximums pushed participants to write even more:

- **175-200 words with GPT**, representing 87.5-100% of the maximum final story word count. 175 words roughly corresponds to the mean word count of solo-writing in Experiment 1 (AI-175), which encourages participants to write slightly more words than what was anticipated for NO-AI.
- **200-250 words with GPT**, representing 100-125% of the maximum final story word count. This condition requires the participant to write more words than the baseline solo-writing condition and the generated story, which may increase feelings of psychological ownership (AI-200).

5.5.2 Results

We use the same analysis as Experiment 1 with detailed statistical test results in [Table 5.5](#).

Table 5.5: Omnibus and post hoc statistical tests for Experiment 2: (a) *Psychological Ownership*, (b) *Duration*, (c) *Word Count*, and (d) *Text Similarity*. Note that *Mental Demand* and *Effort* are excluded as these omnibus tests were not significant.

(a) <i>Psychological Ownership</i>					(b) <i>Duration</i>				
$\chi^2_{3,N=34} = 43.63, p < .001, W = .43$					$\chi^2_{3,N=34} = 11.32, p < .01, W = .11$				
comparisons		p-value		RBC	comparisons		p-value		RBC
AI-150	AI-175	.69	n.s.	.01	AI-150	AI-175	1.0	n.s.	.06
AI-150	AI-200	.69	n.s.	.02	AI-150	AI-200	.03	*	.09
AI-150	NO-AI	< .001	***	.55	AI-150	NO-AI	1.0	n.s.	.04
AI-175	AI-200	.69	n.s.	.04	AI-175	AI-200	.14	n.s.	.02
AI-175	NO-AI	< .001	***	.54	AI-175	NO-AI	1.0	n.s.	.10
AI-200	NO-AI	< .001	***	.55	AI-200	NO-AI	.13	n.s.	.15

(c) <i>Word Count</i>					(d) <i>Text Similarity</i>				
$\chi^2_{3,N=34} = 71.27, p < .001, W = .70$					$\chi^2_{3,N=34} = 61.94, p < .001, W = .61$				
comparisons		p-value		RBC	comparisons		p-value		RBC
AI-150	AI-175	< .001	***	.57	AI-150	AI-175	.46	n.s.	.11
AI-150	AI-200	< .001	***	.99	AI-150	AI-200	.46	n.s.	.12
AI-150	NO-AI	.16	n.s.	.22	AI-150	NO-AI	< .001	***	1.0
AI-175	AI-200	< .001	***	.99	AI-175	AI-200	.77	n.s.	.005
AI-175	NO-AI	.003	**	.38	AI-175	NO-AI	< .001	***	1.0
AI-200	NO-AI	< .001	***	.99	AI-200	NO-AI	< .001	***	1.0

5.5.2.1 Psychological Ownership

Overall, psychological ownership when writing with AI seems to plateau beyond a 150 word minimum (Figure 5.8). There was a significant effect of *CONDITION* on *Psychological Ownership*, and post hoc tests revealed that people felt less *Psychological Ownership* for all conditions that involved AI than NO-AI (MDN = 6.5, IQR = 1.25), and there were no significant differences between AI-150 (MDN = 4.88, IQR = 2.44), AI-175 (MDN = 5.38, IQR = 2.75), and AI-200 (MDN = 4.75, IQR = 2.44).

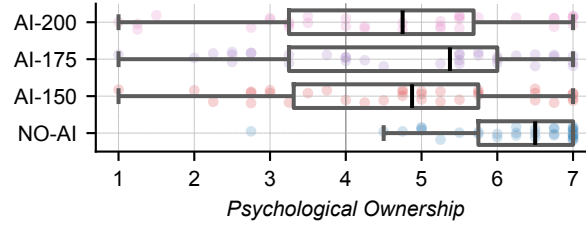


Figure 5.8: *Psychological Ownership* by *CONDITION*. Higher scores correspond to higher feelings of *Psychological Ownership*.

5.5.2.2 Writing Activities

This plateauing effect is further supported by participants' writing activities. Although *Word Count*, *Text Similarity*, *Mental Demand*, and *Effort* were all strongly positively correlated with *Psychological Ownership* in Experiment 1, none of these correlations held in this experiment ($0.22 \leq r_s \leq 0.48$ for significant correlations). Furthermore, there were no meaningful significant effects of *CONDITION* on *Text Similarity*, *Mental Demand*, and *Effort*, despite significant differences in *Word Count*.

5.5.3 Summary

Overall, our results suggest a plateau point around a 150 word minimum for the 150-200 word stories generated in our experiment. Even when participants wrote even longer prompts with AI, they still felt less psychological ownership than writing alone. For this writing task, a prompt length similar to the target word length appears to be a "sweet spot" [134] for trying to maximize psychological ownership when writing with AI, as writing with higher word minimums did not improve psychological ownership.

5.6 ENCOURAGING LONGER PROMPTS IN A REAL INTERFACE

In our experiments, we had to enforce strict prompt length minimums for the purpose of internal validity and experimental control. To give users more flexibility, we explore the use of 'soft' constraints using alternative interface designs and interaction techniques. These are designed to modify the presentation of the output and the prompt entry interface, often in ways

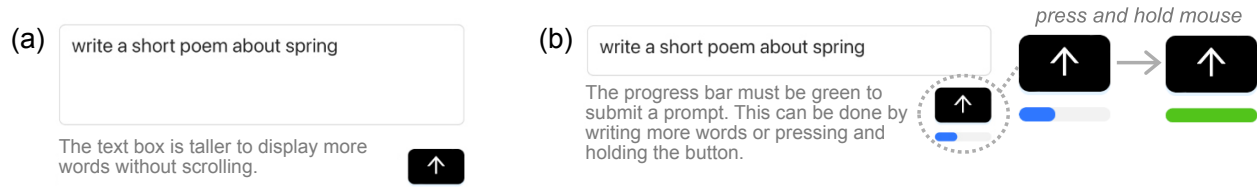


Figure 5.9: Initial prompt entry interface designs that augment the prompt entry area: (a) making the prompt entry text box taller or (b) requiring the user to press and hold the submit button.

that purposely add friction, while remaining fully compatible with current chat-based generative AI tools:

- **Increasing the height of the prompt entry text box** (Figure 5.9a). This approach does not add friction, but could act as a subtle cue for the user to write a longer prompt that fills the height of the text box.
- **Requiring the user to fill a ‘gauge’ by writing more words or pressing and holding the prompt submission button** (Figure 5.9b). These two methods of filling the gauge provide the user with a choice: they can write a very short prompt and press and hold the submission button for a significant amount of time, such as 20 s or more; or they can write a longer prompt and press and hold the button for a shorter amount of time; or write a prompt with a length that meets or exceeds a system-defined “word minimum” and experience no delay at all when submitting.
- **Increasingly blurring the generated output** (Figure 5.10a). If the user writes a shorter prompt, the text is blurred more, and they must press and hold it to make it easier to read. Much like the time delay technique, the blur amount and time spent pressing and holding the output are mapped to prompt length. This technique was inspired by Cockburn et al.’s frost-brushing technique for teaching users about spatial information in user interfaces [44].
- **Increasingly redacting parts of the generated output** (Figure 5.10b). If the user writes a shorter prompt, each chunk of generated output has a higher chance of being redacted. Having a few chunks redacted unlikely hinders overall comprehension as the user can easily ‘fill in the blanks.’ However,

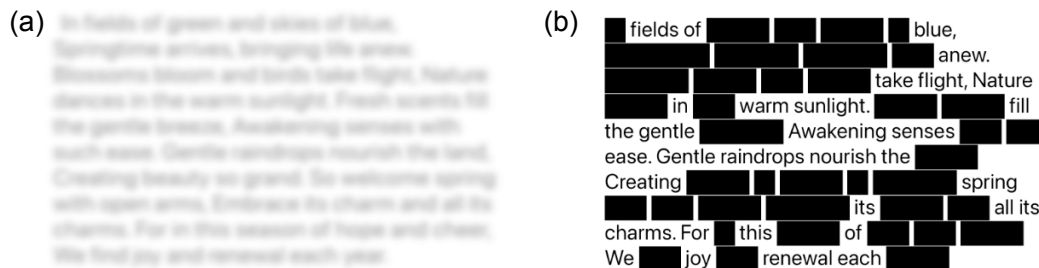


Figure 5.10: Initial prompt entry interface designs that augment the response area: (a) blurring or (b) redacting the text more if the prompt contained less words.

if a majority of the text is redacted and the output is becoming illegible, the user has to increase the length of their existing prompt and resubmit to see more of the generated text.

We ran a pilot experiment (n=90, 11-18 per condition) to better understand which modifications were the most promising for increasing prompt length. Our preliminary results² suggested that integrating a 20-second *time delay* into the submission button could be an effective way to increase prompt length (Figure 5.11).

Given these findings, we investigate the use of time delays in more depth to see whether they could increase prompt length. We suspected that the act of pressing and holding the prompt submission button for several seconds would deter users from writing shorter prompts and encourage users to reflect on their prompts and expand them.

We explain the choice of the 20-second time delay in Section 5.7.4.

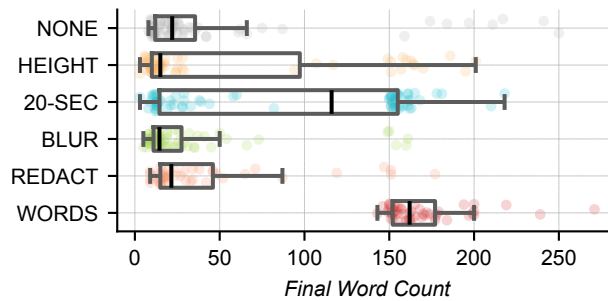


Figure 5.11: *Final Word Count* by *CONDITION* for the initial pilot experiment with all interface modifications.

5.7 EXPERIMENT 3: DELAYED SUBMISSION

The goal of this experiment is to understand the effectiveness of a time delay experienced by pressing and holding the prompt submission button. Participants completed a similar creative writing task as the previous two experiments (writing prompts to generate a 150-200 word short story), but were assigned a random time delay condition for the entire experiment. Two baselines included: no modifications, and a strict word minimum like the previous experiments.

5.7.1 Participants

We recruited 198 new participants with 42 (21%) removed from the analysis for cheating by padding out their responses with dummy characters, copying and pasting a prompt from another source, or experiencing technical issues, leaving 156 valid responses (Table 5.6). All reported being proficient at reading and writing in English (all ≥ 4 on a 1-7 scale). Participants received \$15 as remuneration.

² Data from these participants is also included in the main results of Experiment 3.

Table 5.6: Demographic information for Experiment 3.

Gender		Age		Education		Creative Writing Frequency		Prompt Engineering Familiarity	
Men	93	18-24	5	Less than a High School Diploma	2	Daily	2	Extremely	8
Women	61	25-34	47	High School Diploma	22	Weekly	16	Moderately	26
Non-binary	2	35-44	51	Some University (no credit)	37	Monthly	31	Somewhat	28
		45-54	25	Technical Degree	9	Less than Monthly	73	Slightly	44
		55-64	16	Bachelor's Degree	65	Never	34	Not at All	50
		65-74	11	Professional Degree Beyond Bachelor's Degree	1				
		75+	1	Master's Degree	20				

ChatGPT Frequency	Weekly ChatGPT Use	Other AI Services	AI Usage
Daily 19	0 times 34	Nothing 12 Gemini 65	Nothing 4 Generate Images 68
Weekly 76	1-5 times 70	GPT-2 37 Grammarly 54	Write Emails 52 Get a Definition 59
Monthly 27	5-10 times 32	GPT-3 79 Stable Diffusion 12	Write Papers 21 Explore a Topic 111
Less than Monthly 26	10-30 times 16	GPT-4 60 Dall-E 36	Write Stories 35 Brainstorming 100
Never 8	30+ times 4	Copilot 14 Midjourney 19	Edit Text 65 Find References 35
		Bing 82 Other 13	Generate Code 26 Clarification 67
			Edit Code 19 Translate Text 33
			Debug Code 15 Other 11

5.7.2 Apparatus

The experiment software was similar to that used in the prior experiments, but with a few cosmetic modifications to make the interface more closely resemble existing chat-based AI user interfaces like ChatGPT (Figure 5.12). The prompt entry and output interfaces were stacked on top of each other. The prompt entry text box, which has a default height of 46 pixels, grows by 46 pixels with every new line typed, until a maximum height of 300 pixels.

With the exception of the baseline condition with no modifications, a brief message underneath the prompt entry text box provides instructions for prompt submission (Figure 5.12d). To mitigate risks of bias, the message simply explained how the delayed button worked. Inline with the previous experiments, the time delay mechanics used a 150-word minimum for optimal psychological ownership. The baseline condition had to indicate 150 words as the minimum, but the delayed button conditions made no mention of 150 words.

5.7.2.1 Delayed Button Interaction and Mechanics

The technique is similar to the time delay highlighting technique discussed in Chapter 4, but adjusted to better suit a writing task. A ‘gauge’ that must be filled is visualized as a progress bar below the prompt submission button (Figure 5.12c). When the button is not pressed, the length of the bar is the ratio of the current prompt word count to the *word minimum* (150 words in this experiment). When the button is held down, the progress bar gradually increases (updated every 500 ms) according to an *experienced delay*. This time delay period is proportional to the unfilled bar ratio to the *maximum delay*.

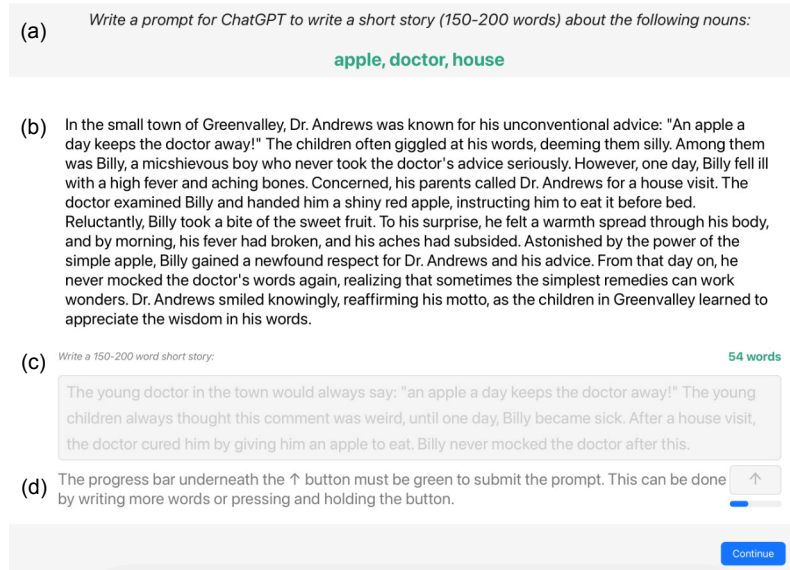


Figure 5.12: Experimental interface with a time delay: (a) top toolbar showing the instructions and three nouns the story had to be about; (b) the story written with AI assistance; (c) a text box where the participant wrote their prompt; (d) instructions about the technique.

For example, consider a 150-word *word minimum* and a *maximum delay* of 20 s. If the user writes a 15-word prompt, the *experienced delay* (the time they must hold the button down) will be 18 s. If the user enters 135 words, the *experienced delay* will only be 2 s. If the user enters at least 150 words, the *experienced delay* will be 0 s (i.e., the button works with a normal click). The progress bar is blue when not full, turning green when enough words have been typed or when enough delay has been experienced by pressing and holding the button. Note that the user must press and hold the button in one single act; releasing it before the progress bar turns green resets the progress bar to its original size based on the number of words in the prompt.

5.7.3 Procedure

The procedure was nearly identical to the prior experiments. However, rather than experiencing all conditions once, participants experienced a single condition four times as we suspected a learning curve for the experimental conditions. As the focus of this study was not psychological ownership, they did not answer questions after every story, instead, they answered questions about their experience after all four trials. The entire experiment took 20 to 60 minutes, depending on the condition.

5.7.4 Choice of Time Delays

Prior work suggests that delays longer than 10 seconds will cause users to shift their attention [125], which is desirable as participants may opt to write

a longer prompt instead of pressing and holding the prompt submission button. Delays of 20 seconds are tolerable [155], so the initial pilot experiment described above tested a 20-second delay. The results suggested a bimodal distribution of prompt word length, participants either wrote very short prompts, or very long prompts (Figure 5.11).

Based on the pilot, we introduce an additional time delay condition: an even longer, 60-second delay [155], to increase frustration and encourage more participants to write longer prompts. We also include a progress bar condition, where the same progress bar visualization was shown underneath the prompt submission button, but the participant did not have to press and hold the button. This was included to ensure that the effects were indeed caused by a time delay, rather than confounding factors such as minor changes to the user interface. Therefore, we test the following five conditions:

- **No Modifications** (n=32), a baseline where participants did not see or experience any modifications to the user interface. This best approximates current chat-based AI interfaces.
- **Progress Bar** (n=32), where the participant sees a progress bar indicating how many words out of 150 they had typed, but do not experience any delay.
- **20-Second Delay** (n=31), where the participant sees the same progress bar indicator and can experience a delay up to 20 seconds long, depending on how many words they write out of 150.
- **60-Second Delay** (n=32), which is similar to the 20-second delay condition, but with a longer, 60-second delay.
- **150-Word Minimum** (n=29), a baseline representing the case where the participant must write a prompt that is at least 150 words long. This condition is guaranteed to increase prompt length, which allows us to make relative comparisons to other conditions.

5.7.5 Design

This is a mixed-design with two primary independent variables: **CONDITION** (levels: NONE, PROGRESS, 20-SEC, 60-SEC, WORDS), which was between-subjects and randomly assigned, and **TRIAL** (levels: 1, 2, 3, 4), which was within-subjects.

The primary measures were objective metrics calculated from logs: the word count of the submitted prompt (*Final Word Count*), the number of times the participant tried to submit a prompt by pressing the prompt submission button (*Submission Attempts*), the word count of the prompt for each submission attempt (*Attempted Word Count*), and the number of seconds they held the button for each submission attempt (*Attempted Duration*). Note that *Submission Attempts*, *Attempted Word Count* and *Attempted Duration* are only applicable for the 20-SEC and 60-SEC conditions, as the prompt submission button could only be pressed once for all other conditions due to the lack of delay (i.e., *Attempted Word Count* is equal to *Final Word Count* for NONE, PROGRESS, and WORDS).

Table 5.7: Omnibus and post hoc statistical tests for Experiment 3: (a) *Final Word Count*, (b) *Submission Attempts*, (c) *Attempted Word Count* for TRIAL = 1.

(a) <i>Final Word Count</i>							
CONDITION ($F_{4,151} = 26.45, p < .001, \eta_p^2 = 0.41$)				TRIAL ($F_{3,453} = 2.87, p < .05, \eta_p^2 = 0.02$)			
comparisons		p-value		comparisons		p-value	
NONE	O-SEC	1.00	n.s.	1	2	.11	n.s.
NONE	20-SEC	.042	*	1	3	.13	n.s.
NONE	60-SEC	.013	*	1	4	.049	*
NONE	WORDS	< .001	***	2	3	1.00	n.s.
O-SEC	20-SEC	.013	*	2	4	1.00	n.s.
O-SEC	60-SEC	.0024	**	3	4	1.00	n.s.
O-SEC	WORDS	< .001	***				
20-SEC	60-SEC	1.00	n.s.				
20-SEC	WORDS	< .001	***				
60-SEC	WORDS	< .001	***				

(b) <i>Submission Attempts</i>				(c) <i>Attempted Word Count</i>			
TRIAL ($F_{3,183} = 96.95, p < .001, \eta_p^2 = 0.61$)				CONDITION ($\chi^2_{4,N=156} = 71.11, p < .001, W = .17$)			
comparisons		p-value		comparisons		p-value	
1	2	< .001	***	NONE	O-SEC	1.00	n.s.
1	3	< .001	***	NONE	20-SEC	1.00	n.s.
1	4	< .001	***	NONE	60-SEC	1.00	n.s.
2	3	.15	n.s.	NONE	WORDS	< .001	***
2	4	.84	n.s.	O-SEC	20-SEC	1.00	n.s.
3	4	.16	n.s.	O-SEC	60-SEC	1.00	n.s.
				O-SEC	WORDS	< .001	***
				20-SEC	60-SEC	1.00	n.s.
				20-SEC	WORDS	< .001	***
				60-SEC	WORDS	< .001	***

5.8 EXPERIMENT 3: RESULTS

Where applicable, we use the Aligned Rank Transform (ART) [181] and post hoc contrast tests (ART-C) [57], Kruskal-Wallis tests and Mann-Whitney U tests, Holm’s corrections for multiple comparisons, and Spearman’s correlations. Statistical test results are in Table 5.7.³

I was unaware that the ART can fail to control for Type I error [172]. But after visualizing the data, I do not suspect this to be an issue here.

³ The ART-C library does not provide effect sizes for post hoc comparisons.

5.8.1 Final Word Count

Overall, experiencing a time delay when submitting a prompt increased prompt length (Figure 5.13). A significant main effect of CONDITION on *Final Word Count* and post hoc tests revealed that 20-SEC (MDN = 31, IQR = 138) and 60-SEC (MDN = 51.1, IQR = 135) was associated with significantly higher *Final Word Count* than NONE (MDN = 18, IQR = 16.25) and PROGRESS (MDN = 18, IQR = 20.5). This was a result of the time delay, rather than changes to the user interface, as there were no significant differences between NONE and PROGRESS. However, as expected, a time delay is not as effective as a strict word minimum, as WORDS had the highest *Final Word Count* (MDN = 157.5, IQR = 17).

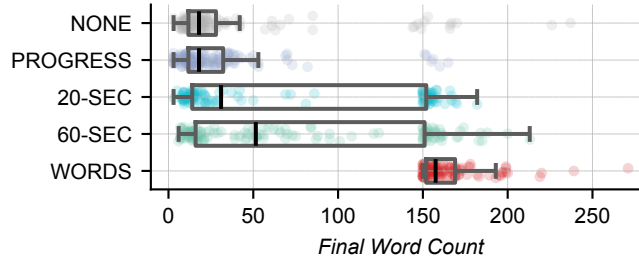


Figure 5.13: *Final Word Count* by CONDITION.

5.8.2 Learning Effect

We expected that participants would learn how to use the delayed prompt submission across multiple trials, but this did not occur. Rather, most participants discovered their preferred way of interacting with the button in the first trial, and continued this for all four trials. Specifically, significant main effects of TRIAL on *Submission Attempts* and *Attempted Word Count* and post hoc tests only revealed significant differences between TRIAL 1 and all other trials. Therefore, we focus our discussion of the learning effect on the results from the first trial.

5.8.2.1 Submission Attempts

Participants made significantly more *Submission Attempts* in the first trial for 20-SEC (MDN = 3, IQR = 2) and 60-SEC (MDN = 3, IQR = 3) than subsequent trials (MDN = 1 and IQR = 0 for all). However, the median *Attempted Duration* was only 0.1 seconds for both 20-SEC and 60-SEC, suggesting that participants were initially trying to press the button normally, without holding it down.

5.8.2.2 Attempted Word Count versus Final Word Count

Examining prompt submission attempts in the first trial, *Attempted Word Count* for 20-SEC (MDN = 17, IQR = 19) and 60-SEC (MDN = 17, IQR = 27) was on par with the *Final Word Count* of NONE (MDN = 15, IQR = 10) and

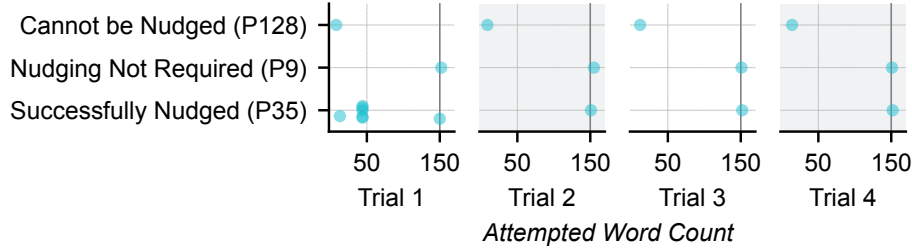


Figure 5.14: Prompt word count each time the submit button was pressed (i.e., *Attempted Word Count*) for these participants. Examples from 20-SEC condition, each *Submission Attempt* is a blue dot.

PROGRESS (MDN = 16, IQR = 14.5), with no significant differences observed. But 20-SEC and 60-SEC were associated with significantly higher *Final Word Count*, suggesting that participants eventually wrote longer prompts after attempting to write shorter prompts. Specifically, in the first trial for 20-SEC and 60-SEC, participants first tried to submit prompts that were 26.81 and 39.16 words long on average, before finally submitting prompts that were 78.10 and 69.03 words long on average, respectively. In other words, *Final Word Count* nearly tripled and doubled in length from the first *Attempted Word Count*.

5.8.3 Strategies

Overall, there were no significant differences between 20-SEC and 60-SEC for *Final Word Count*, and both techniques invoked similar behaviours from participants (Figure 5.13): they either wrote short prompts (≤ 50 words), or long prompts (≥ 150 words). Examining individual prompt submission attempts for the first trial for 20-SEC and 60-SEC provides insights into these diverging strategies. Some participants are consistent and do not change their behaviour, with 29 participants (46%) showing no differences between their first *Attempted Word Count* and *Final Word Count*, but 32 participants (51%) wrote more, with 17 of them (27%) writing at least 50 more words, which consisted of more details about the story plot and writing style. These differences suggest three primary types of users and strategies (Figure 5.14): people who do not need to be nudged as they always write long prompts; people who are more resistant to being nudged and always write shorter prompts; and people who can be nudged to write longer prompts.

Feedback from participants suggested that, as expected, those who do not need to be nudged because they wrote longer prompts did not notice any delay, with comments like “it’s nothing that caught my attention” (P9). Participants who prefer to write shorter prompts noted that pressing and holding the button was a worthwhile trade-off, for example, “it required far less time and effort to hold down the button than to write the stories myself” (P46). This may be related to typing speed. A participant who wrote short prompts said “[someone] who types fast may not mind [writing more]” (P136) while some participants said the opposite, for example: “it was easier to type more into the prompt rather than to hold down the button itself” (P156).

Some felt like they “*hadn’t written enough prompt words*” (P109) when they experienced a delay. For others, the time delay helped them to “*reflect on the prompt*” (P116) and “*make adjustments*” (P105) before it was submitted. This reflection may have encouraged creative thinking, supported by comments like: “*it got me to think more of how I wanted the story to be and pushed me to be more creative*” (P40).

5.8.4 Summary

Overall, our results suggest that a time delay can **nudge** people to write longer prompts. Although participants initially tried to submit shorter prompts in the first trial, they eventually submitted prompts that were significantly longer, and repeated this behaviour for subsequent trials. Some participants cannot or do not need to be nudged, but many who might normally write shorter prompts can be encouraged to write more.

5.9 EXPERIMENT 4: PUMPING A SLIDER

After seeing the effectiveness of a time cost, we were curious to learn if the gauge could be filled in other ways. We propose imposing a *motor cost*, where users must repeatedly move a vertical slider positioned beside the *Submit* button up and down. Like the time cost, this provides users with a choice: they can either write a longer prompt, or exert physical effort to submit a shorter one. Using a similar experimental task, we conducted a within-subjects experiment where participants wrote prompts to generate short stories while experiencing a motor cost, a temporal cost (from Experiment 3), or no interface augmentations.

Table 5.8: Demographic information for Experiment 4.

Gender	Age	Education	Creative Writing Frequency	Prompt Engineering Familiarity
Men	20	18-24 1	High School 3	Weekly 3
Women	9	25-34 5	Some University (no credit) 4	Monthly 8
		35-44 16	Technical Degree 2	Less than Monthly 9
		45-54 4	Bachelor’s Degree 16	Never 9
		55-64 1	Master’s Degree 4	Slightly 9
		65-74 2		Not at All 9

ChatGPT Frequency	Weekly ChatGPT Use	Other AI Services	AI Usage
Daily 11	0 times 3	OpenAI GPT Models 20	Dall-E 3
Weekly 11	1-5 times 8	CoPilot 4	Midjourney 2
Monthly 5	5-10 times 5	Bing 16	Stable Diffusion 1
Less than Monthly 2	10-30 times 6	Gemini 19	Other 4
	30+ times 7	Grammarly 9	
			Write Emails 10
			Get a Definition 19
			Write Papers 4
			Explore a Topic 20
			Write Stories 9
			Brainstorming 20
			Edit Text 12
			Find References 6
			Generate Code 2
			Clarification 18
			Edit Code 2
			Translate Text 9
			Debug Code 1
			Other 1
			Generate Images 20

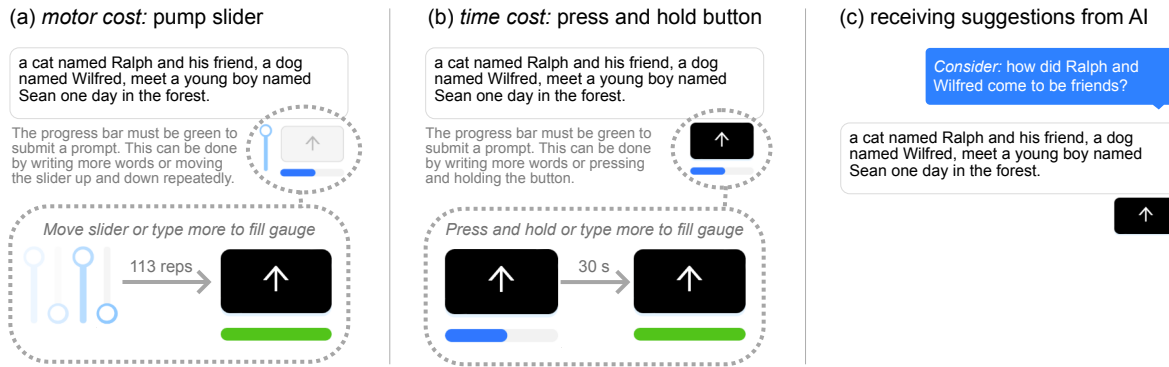


Figure 5.15: Augmented prompt entry techniques that (a) increase the motor cost of submitting a prompt by making the user ‘pump’ a gauge by moving a slider; (b) increase the temporal cost by requiring users to press and hold the button to fill a gauge; and (c) provide AI-generated suggestions for how the user’s prompt can be expanded.

5.9.1 Participants

We recruited 35 new participants from Prolific, using the same inclusion and exclusion criteria as the previous experiments. Six participants (17%) were omitted as they experienced technical issues or who cheated by not making enough keystrokes in the prompt, suggesting they copied and pasted text into the prompt entry text box, leaving 29 valid responses (Table 5.8). All self-reported English reading and writing proficiency (all ≥ 4 on a 7-point scale). Participants received \$7.50 after completing the experiment.

5.9.2 Apparatus

The experiment interface was similar to that of the previous experiment, with a few minor tweaks. First, we tweaked the instructions in the toolbar and above the prompt text box so there were no mentions of the final story length, as we were concerned some participants may mistake this as a minimum prompt length. Second, we used the newer GPT-4.5 Preview model⁴ and fine-tuned the system prompt to generate the responses. Specifically, the system prompts of Experiments 1 and 2 provided example stories to ensure that the response would be the desired length, however, the newer model did not need this. Third, any progress made by pumping the slider, pressing and holding the button, or adding more words to the prompt trigger a 1 s ‘pulsing’ animation to draw the user’s attention to the progress being made. Finally, for the time delay technique (Figure 5.15b), the progress bar expands more frequently (every 100 ms). For the time delay technique, we opted for a 60-second delay, as Experiment 3 suggested that it could encourage users to write longer prompts.

⁴ This model has since been deprecated.

5.9.3 Pumping Slider Interaction and Mechanics

Like the previous experiment, the progress bar represents a ‘gauge’ that the user must fill. This can be done by writing more words in the prompt submission text box, or by ‘pumping’ it full by moving a 45 pixel tall vertical slider up and down repeatedly (Figure 5.15a).

The progress bar width is the ratio of the number of words in the prompt to the *word minimum* (150 words to align with previous experiments). When the user moves the slider up and down, the progress bar width increases. The number of movements the user must perform to fill the gauge is calculated by mapping the ratio of words typed to a *movement maximum*.

We wanted to pick a *movement maximum* that would result in a similar duration as the time delay technique, as differences could have confounded our results. As the time cost technique required users to wait for 60 s in the extreme case, we wanted the extreme case of the motor cost technique to take approximately 60 s as well. To get a sense of how many slider movements would be needed to approximate 60 s, we asked 5 people within our research group to try moving the slider 100 times. The median duration for a single movement was 267 ms. Dividing 60,000 ms by 267 ms means 225 movements to approximate 60 s.

For example, if the user types 15 words, this means the gauge is 10% full. To submit this prompt, the user must then move the slider up and down 203 times (90% of 225). But if the user types 125 words, the gauge is 83% full, so the user only has to move the slider 38 times. If the user types at least 150 words, they do not have to move the slider at all.

Moving the slider up and down to pump up the gauge does not have to be done continuously; the user could lift their mouse after a few movements and continue later. However, when they edit their prompt text, the gauge is reset to reflect the width based on number of words typed. When the gauge is full, the progress bar turns green and submit button is enabled for the user to press without any pumping. Otherwise, the progress bar is blue and the submit button is disabled.

5.9.4 Procedure and Design

The procedure was identical to that of the previous experiments. This is a within-subjects experiment with one primary independent variable, *CONDITION* (levels: *BASELINE*, *TIME*, *MOTOR*). We suspected possible order effects of *CONDITION*, where participants may write more for *BASELINE* after experiencing *TIME* or *MOTOR* first. As such, to get a more accurate measure of *BASELINE*, participants always prompted with this *CONDITION* first. The order of *TIME* and *MOTOR* was randomly assigned. There was only one trial per *CONDITION*. The same measures as Experiments 1 and 2 were collected ($\alpha = .94$ for *Psychological Ownership*), however all subjective measures were on a 0-100 scale. Learning from Experiment 3, we also calculated the *Initial Word Count*, how long the prompt was when the participant first tried to press the button (*TIME*) or move the slider (*MOTOR*).

As mentioned in Chapter 4, I conducted this experiment a bit later than the first three experiments, and thought it would be good to have more granularity in the scores.

Table 5.9: Omnibus and post hoc statistical tests for Experiment 4: (a) *Word Count*, (b) *Text Similarity*, (c) *Psychological Ownership*, (d) *Mental Demand*, (e) *Physical Demand*, (f) *Effort*, and (g) *Frustration*.

(a) <i>Word Count</i>					(b) <i>Text Similarity</i>				
$\chi^2_{2,N=29} = 16.72, p < .001, W = .29$					$\chi^2_{2,N=29} = 7.52, p < .05, W = .13$				
comparisons		p-value		RBC	comparisons		p-value		RBC
BASELINE	TIME	< .001	***	.45	BASELINE	TIME	.03	*	.38
BASELINE	MOTOR	< .001	***	.54	BASELINE	MOTOR	.01	*	.38
TIME	MOTOR	.37	n.s.	.10	TIME	MOTOR	.67	n.s.	.05

(c) <i>Psychological Ownership</i>					(d) <i>Mental Demand</i>				
$\chi^2_{2,N=29} = 9.95, p < .01, W = .13$					$\chi^2_{2,N=29} = 15.21, p < .001, W = .26$				
comparisons		p-value		RBC	comparisons		p-value		RBC
BASELINE	TIME	.04	*	.26	BASELINE	TIME	.006	**	.45
BASELINE	MOTOR	.04	*	.29	BASELINE	MOTOR	< .001	***	.50
TIME	MOTOR	.47	n.s.	.05	TIME	MOTOR	.92	n.s.	.01

(e) <i>Physical Demand</i>					(f) <i>Effort</i>					(g) <i>Frustration</i>				
$\chi^2_{2,N=29} = 37.90, p < .001, W = .65$					$\chi^2_{2,N=29} = 23.79, p < .001, W = .41$					$\chi^2_{2,N=29} = 16.53, p < .001, W = .28$				
comparisons		p-value		RBC	comparisons		p-value		RBC	comparisons		p-value		RBC
BASELINE	TIME	< .001	***	.52	BASELINE	TIME	< .001	***	.62	BASELINE	TIME	< .001	***	.52
BASELINE	MOTOR	< .001	***	.68	BASELINE	TIME	< .001	***	.59	BASELINE	MOTOR	.01	*	.49
TIME	MOTOR	.02	*	.25	TIME	MOTOR	.19	n.s.	.07	TIME	MOTOR	.57	n.s.	.06

5.9.5 Results

We use the same statistical tests as Experiments 1 and 2, with detailed test results shown in [Table 5.9](#).

5.9.5.1 Word Count

Overall, both interaction techniques increased prompt length ([Figure 5.16](#)). There was a significant effect of CONDITION on *Word Count*, with post hoc tests revealing that TIME (MDN = 35, IQR = 80) and MOTOR (MDN = 49, IQR = 91) were both associated with longer prompts than BASELINE (MDN = 15, IQR = 23). *Word Count* more than doubled for TIME and tripled for MOTOR. Seven participants (24%) noted how the interaction techniques “gave some time to reflect” (P14), and encouraged them to “write more and add more detail” (P22).

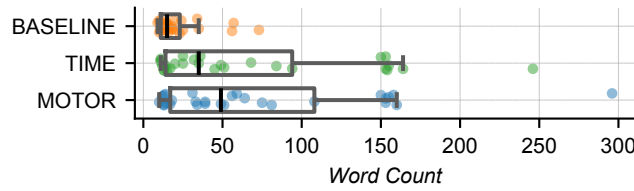


Figure 5.16: Word Count by CONDITION.

5.9.5.2 Text Similarity

Both interaction techniques had higher similarities between the prompt and output. A significant effect of CONDITION on *Text Similarity* and post hoc tests revealed that the generated stories for TIME (MDN = .37, IQR = .18) and MOTOR (MDN = .36, IQR = .12) shared more similarities with the prompts than BASELINE (MDN = .28, IQR = .15). *Text Similarity* was moderately positively correlated with *Word Count* ($r_s = .48, p < .0001$), suggesting that as participants wrote longer prompts, the resulting stories more closely resembled them.

5.9.5.3 Nudging Mechanics

Although there were significant improvements in *Word Count* overall, TIME and MOTOR did not work for all participants. For TIME, participants tried to push the button 10 times on average. When comparing participants' *Initial Word Count* to their final *Word Count*, participants wrote significantly fewer words initially ($p < .01$; MDN = 20, IQR = 36). Twelve (41%) increased their *Word Count* (MDN = 42.5, IQR = 81.5 for differences), with 4 (14%) increasing by 50 words or more. However, majority (16, 55%) showed no differences between their *Initial Word Count* and final *Word Count*, and of these, only 1 wrote a prompt that was at least 150 words. These results are similar to Experiment 3.

We observed a similar trend with MOTOR. Participants tried to move the slider 6 times on average. Participants wrote significantly fewer words initially ($p < .01$; MDN = 31, IQR = 28). Twelve (41%) increased their *Word Count* from their *Initial Word Count* (MDN = 31.5, IQR = 52.25 for differences), with 5 (17%) increasing by 50 words or more. Majority of participants (16, 55%) had no difference, and of these, 5 (17%) already wrote prompts that were at least 150 words long, meaning they did not have to move the slider at all.

5.9.5.4 Psychological Ownership

Both interaction techniques were associated with higher feelings of psychological ownership (Figure 5.17). A significant effect of CONDITION on *Psychological Ownership* with post hoc tests revealed that TIME (MDN = 41, IQR = 54) and MOTOR (MDN = 47.5, IQR = 57.25) had higher *Psychological Ownership* than BASELINE (MDN = 28.25, IQR = 39), which was supported by comments like: "I feel [like the time delay] was prompting me to be more involved in the story details. I would spend more time on the prompt itself. It made the story a bit more personal" (P15).

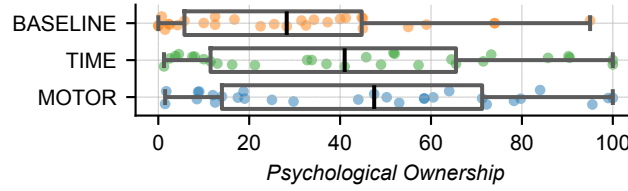


Figure 5.17: *Psychological Ownership* by CONDITION. Higher scores correspond to higher feelings of psychological ownership.

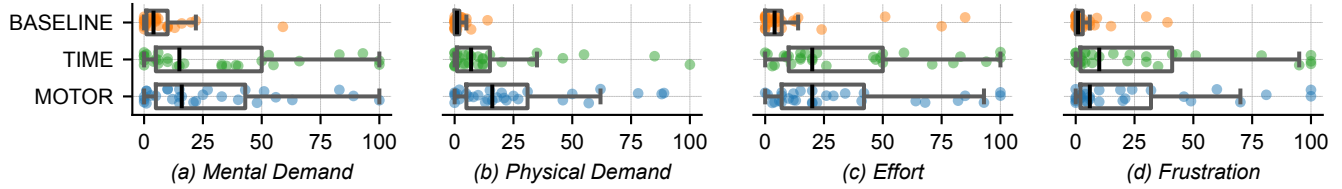


Figure 5.18: *Mental Demand*, *Physical Demand*, *Frustration*, and *Effort* by CONDITION.

Although participants wrote longer prompts with TIME and MOTOR, *Word Count* and *Text Similarity* were only weakly positively correlated with *Psychological Ownership* (both $r_s = .28, p < .01$). This may suggest that workload-related factors of the techniques contributed more to the higher feelings of psychological ownership than just prompt length and similarity.

5.9.5.5 Task Workload

Both interaction techniques increased *Mental Demand*, *Physical Demand*, *Effort*, and *Frustration* (Figure 5.18). For *Mental Demand*, scores were lower for BASELINE (MDN = 4, IQR = 9) than TIME (MDN = 15, IQR = 45) and MOTOR (MDN = 16, IQR = 38). *Mental Demand* was moderately positively correlated with *Word Count* ($r_s = .48, p < .0001$) and *Psychological Ownership* ($r_s = .5, p < .0001$). MOTOR had the highest *Physical Demand* (MDN = 16, IQR = 26), followed by TIME (MDN = 7, IQR = 14), and BASELINE (MDN = 1, IQR = 2). For *Effort*, BASELINE (MDN = 4, IQR = 7) resulted in lower scores than TIME (MDN = 20, IQR = 40) and MOTOR (MDN = 20, IQR = 35). *Effort* was weakly positively correlated with *Word Count* ($r_s = .35, p < .001$) and moderately positively correlated with *Psychological Ownership* ($r_s = .42, p < .0001$). BASELINE (MDN = 1, IQR = 3) had lower *Frustration* than TIME (MDN = 10, IQR = 39) and MOTOR (MDN = 6, IQR = 30). No other metrics or correlations were significant.

These findings suggest that the increased *Mental Demand* and *Effort* required to submit prompts with TIME and MOTOR contributed to higher feelings of *Psychological Ownership*. This was supported by comments like: “I think [the slider] encourages me to add more of a personal input in the story that is submitted. Although it was a bit annoying, the end result felt more of a personal effort than not” (P20); and “having to move the slider handle up and down felt like work because it’s something I had to do in order to submit the prompt. It made me feel a slightly deeper connection since I sort of had to work for it” (P18).

5.9.6 Summary

Overall, our results suggest that increasing the temporal and motor cost of submitting a prompt can encourage users to write longer prompts. These techniques required additional mental and physical demand, and were more effortful and frustrating. However, the increased prompt length, text similarity, mental demand, and effort of these techniques all contributed to higher feelings of psychological ownership. These findings validate and significantly extend those of Experiment 3, and provide evidence that incorporating motor costs into the prompt submission process can improve psychological ownership.

5.10 RECEIVING SUGGESTIONS FROM AI

Encouraged by these results, we were interested in further augmenting these techniques to see if they could further improve psychological ownership. Prior work shows that receiving suggestions from AI can lower [51, 103] or increase [104] psychological ownership depending on the interaction. Even in writing scenarios that do not involve AI, receiving feedback can help writers [99] and suggestions can encourage ideation and creativity, and help to overcome writer’s block [160], which may occur when writing longer prompts. Receiving suggestions while waiting or moving the slider may further nudge users to write longer prompts, which may further increase mental demand, effort, and psychological ownership.

We propose augmenting these techniques to display suggestions for how the prompt can be expanded when the user stops writing. After the user stops writing for 1 second, a blue tooltip appears on the top right corner of the prompt submission box (Figure 5.15c). This resembles the “on-idle” condition from Siangliulue et al. [160]. The tooltip contains a suggestion, phrased as a question to encourage more thought [49], that is one sentence long for how the prompt can be expanded or improved to incorporate more details (e.g., *Consider: how does the relationship between the characters influence the story?*) These suggestions are generated using GPT 3.5 Turbo for increased speed. Importantly, the suggestion is not automatically incorporated as edits to the prompt as this may reduce psychological ownership [104]. The tooltip stays on the screen at maximum opacity for 4 seconds, then gradually fades away for 3 seconds, mimicking the mechanics of prior work I conducted on displaying notifications [83]. This gives users time to choose whether they would like to incorporate these suggestions or not. If the user begins typing again, the tooltip does not show up (unless they again stop typing for 1 second). Otherwise, the tooltip reappears after 1 second.

5.11 EXPERIMENT 5: PROMPT SUGGESTIONS

We recruited 30 new participants using the same inclusion and exclusion criteria as before. No participants had to be filtered out (Table 5.10). All self-reported English reading and writing proficiency (all ≥ 4 on a 7-point scale). The task and procedure were both identical to that of Experiment 4. The ap-

Table 5.10: Demographic information for Experiment 5.

Gender		Age		Education		Creative Writing Frequency		Prompt Engineering Familiarity	
Men	15	18-24	2	High School	4	Daily	1	Extremely	3
Women	15	25-34	6	Some University (no credit)	2	Weekly	6	Moderately	6
		35-44	6	Technical Degree	2	Monthly	6	Somewhat	5
		45-54	10	Bachelor's Degree	17	Less than Monthly	12	Slightly	7
		55-64	4	Master's Degree	4	Never	5	Not at All	9
		65-74	2	Doctorate	1				

ChatGPT Frequency		Weekly ChatGPT Use		Other AI Services				AI Usage			
Daily	8	0 times	4	Nothing	2	Grammarly	12	Nothing	2	Generate Images	15
Weekly	12	1-5 times	12	OpenAI GPT Models	23	Dall-E	5	Write Emails	14	Get a Definition	17
Monthly	5	5-10 times	5	CoPilot	4	Midjourney	4	Write Papers	7	Explore a Topic	17
Less than Monthly	3	10-30 times	6	Bing	12	Stable Diffusion	3	Write Stories	9	Brainstorming	20
Never	2	30+ times	3	Gemini	18	Other	3	Edit Text	17	Find References	9
								Generate Code	5	Clarification	13
								Edit Code	4	Translate Text	9
								Debug Code	3	Other	1

paratus was mostly the same, however, each technique was further enhanced to display suggestions (Figure 5.15c). These suggestions were integrated into the same interaction techniques and baseline as before. For all conditions, the prompt submission button was disabled until at least one suggestion was shown. The design was mostly identical to that of Experiment 4, however, we also considered *Number of Suggestions*, the number of suggestions each participant saw for a single CONDITION.

5.11.1 Results

We use the same statistical tests as Experiment 4, with detailed statistical test results in Table 5.11.

5.11.1.1 Word Count

Both interaction techniques increased prompt length (Figure 5.19). A significant effect of CONDITION on *Word Count* and post hoc tests showed that BASELINE (MDN = 19, IQR = 45.25) resulted in shorter prompts than TIME (MDN = 155, IQR = 43) and MOTOR (MDN = 150.5, IQR = 64.75). This represents a difference of over 100 words.

5.11.1.2 Text Similarity

Both interaction techniques also increased text similarity. A significant effect of *Text Similarity* on CONDITION revealed that prompts written with BASELINE (MDN = .32, IQR = .14) were less similar than TIME (MDN = .41, IQR = .13) and MOTOR (MDN = .42, IQR = .14). *Text Similarity* was moderately positively correlated with *Word Count* ($r_s = .55, p < .0001$).

Table 5.11: Omnibus and post hoc statistical tests for Experiment 5: (a) *Word Count*, (b) *Text Similarity*, (c) *Psychological Ownership*, (d) *Mental Demand*, (e) *Physical Demand*, (f) *Effort*, (g) *Frustration*, and (h) *Number of Suggestions*.

(a) <i>Word Count</i>					(b) <i>Text Similarity</i>				
$\chi^2_{2,N=30} = 35.06, p < .001, W = .58$					$\chi^2_{2,N=30} = 14.6, p < .001, W = .24$				
comparisons		p-value		RBC	comparisons		p-value		RBC
BASELINE	TIME	< .001	***	.70	BASELINE	TIME	< .001	***	.50
BASELINE	MOTOR	< .001	***	.67	BASELINE	MOTOR	.002	**	.45
TIME	MOTOR	.83	n.s.	.11	TIME	MOTOR	.79	n.s.	.02

(c) <i>Psychological Ownership</i>					(d) <i>Mental Demand</i>				
$\chi^2_{2,N=30} = 14.36, p < .001, W = .24$					$\chi^2_{2,N=30} = 10.14, p < .01, W = .17$				
comparisons		p-value		RBC	comparisons		p-value		RBC
BASELINE	TIME	< .001	***	.35	BASELINE	TIME	.003	**	.44
BASELINE	MOTOR	.005	**	.23	BASELINE	MOTOR	.007	**	.41
TIME	MOTOR	.02	*	.15	TIME	MOTOR	.98	n.s.	.04

(e) <i>Physical Demand</i>					(f) <i>Effort</i>				
$\chi^2_{2,N=30} = 10.82, p < .01, W = .18$					$\chi^2_{2,N=30} = 23.54, p < .001, W = .39$				
comparisons		p-value		RBC	comparisons		p-value		RBC
BASELINE	TIME	.04	*	.24	BASELINE	TIME	< .001	***	.56
BASELINE	MOTOR	.007	**	.31	BASELINE	MOTOR	< .001	***	.61
TIME	MOTOR	.14	n.s.	.08	TIME	MOTOR	.74	n.s.	.02

(g) <i>Frustration</i>					(h) <i>Number of Suggestions</i>				
$\chi^2_{2,N=30} = 8.90, p < .05, W = .15$					$\chi^2_{2,N=30} = 41.30, p < .001, W = .69$				
comparisons		p-value		RBC	comparisons		p-value		RBC
BASELINE	TIME	.05	n.s.	.31	BASELINE	TIME	< .001	***	.82
BASELINE	MOTOR	.003	**	.50	BASELINE	MOTOR	< .001	***	.79
TIME	MOTOR	.05	n.s.	.22	TIME	MOTOR	.85	n.s.	.009

5.11.1.3 *Nudging Mechanics*

For TIME, participants pressed the button 7 times on average. When comparing the *Initial Word Count* to the final *Word Count*, participants wrote significantly fewer words initially ($p < .001$; MDN = 91.5, IQR = 126.5). Eleven (37%) were consistent, of which 7 (23%) consistently wrote at least 150 words. Majority of participants (18, 60%) wrote more (MDN = 63, IQR = 110), with 11 (37%) writing 50 words or more.

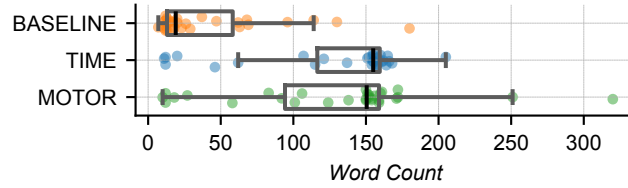


Figure 5.19: *Word Count* by *CONDITION* when receiving AI-generated suggestions.

For *MOTOR*, participants moved the slider 6 times on average. Again, participants wrote significantly fewer words initially ($p < .01$; $MDN = 96.5$, $IQR = 134.75$). Majority (20, 67%) were consistent, with 13 (43%) writing at least 150 words and not having to interact with the slider at all. Ten (33%) wrote more ($MDN = 85$, $IQR = 53.5$ for differences), with 8 (27%) writing at least 50 words.

5.11.1.4 *Psychological Ownership*

Both interaction techniques were associated with higher feelings of psychological ownership (Figure 5.20). A significant effect of *CONDITION* on *Psychological Ownership* and post hoc tests showed that *TIME* ($MDN = 59.12$, $IQR = 39.38$) had the highest *Psychological Ownership*, followed by *MOTOR* ($MDN = 49.38$, $IQR = 49.31$), and *BASELINE* ($MDN = 39.5$, $IQR = 44$). *Psychological Ownership* was weakly positively correlated with *Word Count* and ($r_s = .37$, $p < .001$) and *Text Similarity* ($r_s = .32$, $p < .01$).

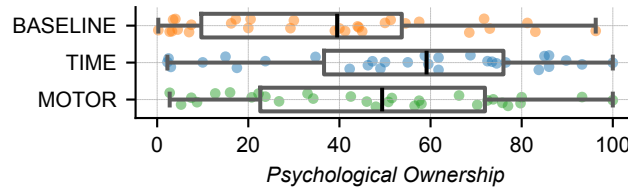


Figure 5.20: *Psychological Ownership* by *CONDITION* when receiving AI-generated suggestions.

5.11.1.5 *Task Workload*

Both interaction techniques increased *Mental Demand*, *Physical Demand*, *Effort*, and *Frustration* (Figure 5.21). For *Mental Demand*, *BASELINE* ($MDN = 10$, $IQR = 23.5$) resulted in lower scores than both *TIME* ($MDN = 36$, $IQR = 37.75$) and *MOTOR* ($MDN = 30.5$, $IQR = 52.75$). *Mental Demand* was moderately positively correlated with *Word Count* and strongly positively correlated with *Psychological Ownership* ($r_s = .62$, $p < .0001$). *BASELINE* ($MDN = 3$, $IQR = 8.75$) required less *Physical Demand* than *TIME* ($MDN = 7.5$, $IQR = 31.25$) and *MOTOR* ($MDN = 9$, $IQR = 36.5$). There was a weak positive correlation between *Physical Demand* and *Psychological Ownership* ($r_s = .35$, $p < .001$). For *Effort*,

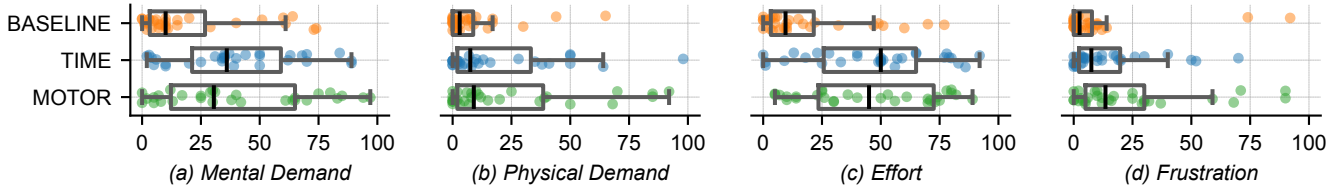


Figure 5.21: Mental Demand, Physical Demand, Frustration, and Effort by CONDITION when receiving AI-generated suggestions.

BASELINE (MDN = 9.5, IQR = 18.25) was less demanding than TIME (MDN = 50, IQR = 39.25) and MOTOR (MDN = 45, IQR = 49). Effort was moderately positively correlated with Word Count ($r_s = .43, p < .0001$) and Psychological Ownership ($r_s = .5, p < .0001$). MOTOR (MDN = 13.5, IQR = 25) resulted in higher Frustration than BASELINE (MDN = 2.5, IQR = 7.75). No other metrics or correlations were statistically significant.

5.11.1.6 Number of Suggestions

Participants received more suggestions when it was integrated into the two interaction techniques. A significant effect of CONDITION on Number of Suggestions and post hoc tests revealed that participants received fewer suggestions with BASELINE (MDN = 4.5, IQR = 8.25) than TIME (MDN = 22, IQR = 12) and MOTOR (MDN = 23, IQR = 14.5). Number of Suggestions was moderately positively correlated with Word Count ($r_s = .55, p < .0001$), Text Similarity ($r_s = .45, p < .0001$), Mental Demand ($r_s = .49, p < .0001$), and Effort ($r_s = .54, p < .0001$); and weakly positively correlated with Physical Demand ($r_s = .22, p < .05$), Frustration ($r_s = .25, p < .05$), and Psychological Ownership ($r_s = .26, p < .05$).

Although four participants (13%) noted that the suggestions were “distracting or annoying” (P17), 24 (80%) appreciated them, as they helped when participants “started to feel a bit stuck” (P10) and encouraged more thought, for example: “[the suggestions] got me really thinking and making this more of a collaborative process than just entering a prompt and taking what the AI generated” (P28).

5.11.2 The Effects of Receiving Suggestions

We use Mann-Whitney U tests to make between-subjects comparisons with Experiment 4 to examine the effect of receiving AI-generated suggestions (levels: NONE, from Experiment 4; and SUGGESTIONS, from Experiment 5). Statistical test details are in Table 5.12.

Overall, participants wrote over 100 more words with both interaction techniques when they received suggestions. For both TIME and MOTOR, Word Count was significantly higher with SUGGESTIONS, but receiving suggestions did not affect BASELINE. We observed a similar trend for Text Similarity, where SUGGESTIONS increased similarity for TIME and MOTOR, but not BASELINE. For TIME, participants made their first submission attempt when Initial Word Count was significantly longer with SUGGESTIONS, but this did not occur for MOTOR.

Table 5.12: Between-subjects statistical tests comparing NONE and SUGGESTIONS for: (a) BASELINE, (b) TIME, and (c) MOTOR.

metric	(a) BASELINE			(b) TIME			(c) MOTOR		
	p-value		RBC	p-value		RBC	p-value		RBC
Word Count	.10	n.s.	.25	< .001	***	.55	.003	**	.44
Text Similarity	.22	n.s.	.19	.04	*	.31	.02	*	.37
Initial Word Count	n/a	n/a	n/a	.01	*	.31	.44	n.s.	.14
Psychological Ownership	.25	n.s.	.17	.16	n.s.	.16	.73	n.s.	.05
Mental Demand	.008	**	.40	.10	n.s.	.25	.11	n.s.	.24
Physical Demand	.03	*	.31	.92	n.s.	.02	.46	n.s.	.11
Temporal Demand	.01	*	.32	.22	n.s.	.18	.20	n.s.	.19
Effort	.04	*	.31	.10	n.s.	.25	.07	n.s.	.28

This could explain why there were significant differences in *Psychological Ownership* between TIME and MOTOR for SUGGESTIONS.

Even though participants wrote significantly more words and had more similar texts with TIME and MOTOR, this did not further improve *Psychological Ownership* as no comparisons with NONE were significant. For BASELINE, having SUGGESTIONS increased *Mental Demand*, *Physical Demand*, *Temporal Demand*, and *Effort*, but there were no differences for TIME and MOTOR.

5.11.3 Summary

Overall, our results show that providing AI-generated suggestions can help users write even longer prompts when using techniques that increase the temporal and motor costs of submitting prompts. However, this did not further improve psychological ownership when compared to providing no suggestions.

5.12 DISCUSSION

Our first two experiments show that writing with a chat-based generative AI assistant lowers feelings of psychological ownership. These results extend and support prior work studying psychological ownership when writing with other human collaborators and with AI. The most profound finding from our work is a strong indication that when writing with a chat-based generative AI assistant, psychological ownership can be improved by requiring or encouraging users to write longer prompts. Specifically, writing longer prompts encouraged users to provide more details about the story plot, exerting more control over the output and making it more similar to their original prompt. This often required more mental demand and effort, which are important contributing factors to psychological ownership.

To encourage users to write longer prompts for increased psychological ownership, we propose two simple changes to the prompt submission button: the user must press and hold the button for a delay time or move a slider up

and down if their prompt is too short. Our results show that these techniques typically required more mental demand and effort, but critically, they also increased psychological ownership. Although incorporating AI-generated suggestions into the prompt entry interface further increased prompt length, it did not further improve psychological ownership.

These findings from Experiment 5 expand and provide additional nuance to Experiment 1 and 2. One possibility is that factors beyond prompt length and text similarity are critical. It could have been that workload-related factors were the most important contributors. Unlike word count and text similarity, which were only weakly positively correlated with psychological ownership, mental demand and effort were notably both moderately positively correlated with psychological ownership in Experiment 4 and 5. For an additional check, we conducted a variable importance analysis using the random forest method, and found that mental demand was the most important contributor to increased psychological ownership for both of these experiments.

This is not to say that encouraging users to write prompt length is not a valuable outcome; rather, encouraging users to write longer prompts is one way to encourage more thought and improve psychological ownership.

5.12.1 *Applicability to Other Contexts*

Our experimental setup focused on a creative writing context, but our findings and interaction techniques could be adapted to suit other writing contexts and even be beneficial for other outcomes beyond psychological ownership.

5.12.1.1 *Other Writing Contexts*

People may feel differently about owning different types of text. For example, Nicholes [124] found that undergraduate students feel more psychological ownership towards creative writing while graduate students feel more psychological ownership towards academic writing. Writers may also have different priorities and goals, with some valuing productivity and financial gain over creative fulfillment [15, 108]. Generative AI writing assistants can be used as a tool to optimize for productivity, so depending on the goal, reduced psychological ownership may be a worthwhile trade-off for writers. Repeating a similar experiment but for a wider range of document types and writer goals would give insights into these trade-offs. In cases where psychological ownership is less important to writers, shorter prompts may suffice. One possibility is to further modify the interface to elicit goals from writers [99] before using the generative AI assistant, and use these goals to determine whether the system should invoke techniques to encourage longer prompts from the writer. As the writer's goals change, such techniques could dynamically adapt to better suit them. For example, if the user's goal is to write an abstract, the system could require a longer prompt, but this requirement could be lifted or adjusted if the user's goal changes to receive edits for a specific sentence.

5.12.1.2 *Beyond Psychological Ownership*

Encouraging or requiring longer prompts could be beneficial to encourage other positive outcomes for users. For example, there are growing concerns that “effortlessly generated information” provided by generative AI assistants will reduce critical thinking and problem-solving skills among students [86]. Encouraging or enforcing longer prompts that increase the mental demand and effort could potentially mitigate such risks. Specifically, requiring students to write longer prompts could encourage them to put more thought and effort into core learning outcomes before receiving output from a generative AI assistant (e.g., thinking about the arguments and structure of an essay, or thinking about a coding algorithm). Studying the effect of prompt length on other outcomes, like student learning outcomes, is an interesting avenue for future work.

5.12.2 *Implementation Approaches and Challenges*

We show that longer prompts can be encouraged through simple user interface adjustments: adding a submission button time delay and repeatedly moving a slider. However, there are related design considerations and possibilities for other approaches.

5.12.2.1 *Identifying Optimal “Word Minimums”*

Our piloted interaction techniques all attempted to nudge users towards a known word minimum of 150 words as our first two experiments suggested that psychological ownership plateaued at 150 words for the given task. This 150-word minimum likely does not apply to other types of tasks and user goals, necessitating additional research and methods for identifying prompt lengths that optimize for psychological ownership. Another option is to create interaction techniques and interface augmentations that do not require any knowledge of optimal word minimums. For example, as the user writes, the generative AI assistant could dynamically append a sentence fragment to the prompt to elicit more information (e.g., “Here is some more background information:”). This approach is similar to receiving AI-generated suggestions, but becomes mandatory and automatically integrated into the prompt.

5.12.2.2 *Other Nudging Techniques*

We piloted several modifications to the user interface (Section 5.6) before investigating the use of time and motor costs to nudge users to write longer prompts. These approaches requires minimal changes and can easily be integrated into current chat-based generative AI interfaces. However, other approaches may also encourage longer prompts and may be more effective than time delays. For example, the prompt entry interface could enforce a different prompt structure through scaffolding techniques [113] (e.g., by breaking the prompt into multiple but smaller responses or requiring the user to “fill-in-the-blanks” of a pre-defined prompt).

We investigate some of these ideas in the context of AI margin notes in Chapter 6.

5.12.2.3 *Non-Chat-Based Interfaces*

Encouraging users to write longer prompts is immediately relevant to chat-based generative AI services like ChatGPT, but as AI becomes more integrated through techniques like direct manipulation [118], factors other than prompt length may become more dominant. For example, sketches convey more information than text [74], and sketching can be an effective prompting technique [188]. Future work could explore the role of sketching to elicit more details from users for increased psychological ownership, which could benefit writing and even image generation tasks. An important implication of our work is that designers should think of ways to encourage users to invest more when interacting with generative AI assistants for increased psychological ownership.

5.12.3 *Limitations*

We chose to explore the role of prompt length within a single prompt, where the user mimics the role of a writing “consultant” [144]. In practice, users likely refine generated text over multiple prompts. A natural extension of our work is to explore the effect of word count across multiple messages within a conversation. This is representative of a broader class of user interactions with systems like ChatGPT, where users iterate over the output through multiple prompts.

One possible reason for the plateauing psychological ownership identified in Experiment 2 may be plateauing semantic text similarity. As models continue to improve to incorporate all details specified in prompts, this plateauing effect may disappear, necessitating further investigations of this effect for newer models. We did test if newer GPT-4 Turbo and 4-o models generated more similar output using the same prompts entered by our participants, but we did not observe any differences in *Text Similarity* between the older and newer models.

Another approach to improve psychological ownership is to design systems that incorporate all details indicated by users and that encourage humans to exert more control over AI-generated output (e.g., by manually accepting suggestions [104]). We chose to implement and investigate systems that work like existing, widely-used commercial tools such as ChatGPT for increased ecological validity. However, as other generative AI systems that are designed to increase psychological ownership become more common, our results could serve as a valuable baseline for future studies, or could even be integrated into new systems for even higher feelings of psychological ownership.

The results from Experiment 3 may be lacking in ecological validity. Our experiment did not provide additional rewards to participants who wrote longer prompts, however, the effects of a time delay on prompt length may have been more pronounced due to the use of crowdworkers on Prolific, who are known to be highly conscientious participants [52]. As such, they may have been more willing to write longer prompts to do the task “well” and less likely to abandon using such systems than the general public. Similarly, the 20- and 60-second time delays worked well in the context of this experiment, but they may not transfer well to real-world systems. It is possible that

even shorter time delays, such as 2-second delays studied in prior work [106, 122, 159], may also increase prompt length while mitigating risks of users abandoning the system [97]. However, when we explored the use of shorter time delays in Chapter 4, we did not observe differences in behaviour. Providing users with feedback and detailed explanations on why a delay exists can also increase their willingness to wait [56, 122]. Deploying variations of this nudging technique in real-world systems would give more insights into optimal design parameters that strike a balance between increasing prompt length and preventing system abandonment.

Our results from Experiment 5 experiment may be limited by how we chose to display the prompt suggestions. Given the short writing task and the design of the two interaction techniques we tested, we felt that showing suggestions after 1 s of inactivity was appropriate, however, it may have been more appropriate to display them after a longer period or to display them on-demand (e.g., after the participant presses a button [160]). On-demand suggestions may also result in increased psychological ownership due to the explicit action required from the user to receive a suggestion [104].

5.13 CONCLUSION

Our work shows that writing with an AI assistant reduces feelings of psychological ownership, but requiring or encouraging users to write longer prompts can improve it. This is no replacement for writing something independently, but it is a simple way to encourage writers to include more details about the content and put more thought and effort into their prompts. Experiencing a time delay and moving a slider when submitting a short prompt can effectively increase prompt length along with psychological ownership, simple interface changes to positively impact human-AI collaborations.

INTEGRATING PROMPTING WITH DOCUMENT TEXT THROUGH AI MARGIN NOTES

In the previous chapter, we prevented prompt submission until the user wrote enough words in the prompt submission box of a traditional chat-based generative AI interface. In this chapter, we propose integrating prompting into the commenting feature of document reader software. Unlike chat-based prompting, which makes no assumptions about location, our approach requires users to specify *where* prompts and responses should go by selecting document text, thereby constraining the location. We also propose different types of commenting techniques that involve the human and the AI in different ways, and many techniques require specific action from the user before they could continue, restricting *when* they could finalize their comments. To explore these ideas, we return to an [active reading](#) task that involved prompting to write notes, effectively combining elements of both reading and writing.

6.1 MOTIVATION

Taking notes while reading documents is a common [active reading](#) strategy that can be done in two primary ways. First, people can write notes that are decoupled from the document text by writing on a separate piece of paper. Second, people can write notes that are integrated with the document text [\[2\]](#), for example, by writing in the margins (*marginalia*). Such ‘margin notes’ have been common practice for centuries, as they provide a space for readers to summarize, paraphrase, explain unfamiliar concepts, make connections to existing knowledge, and even express personal opinions [\[80, 143\]](#). This can be especially beneficial when reading for educational purposes [\[145\]](#), as it provides space for people to work through difficult sections and quickly re-access their thoughts afterwards within a shared context [\[2, 115\]](#). When reading digital documents, margin notes can be created using pen input devices (e.g., [\[150\]](#)), but a more common way is by leaving *comments* in systems like Google Docs, Microsoft Word, and Adobe Acrobat Reader [\[184\]](#). This is usually done by selecting text within a document and writing in a text box beside it.

More recently, people have been leveraging the capabilities of LLMs to support reading documents, for example, to summarize, reword, and better understand unfamiliar concepts [\[68, 69, 98\]](#). The generated output serves a purpose like margin notes, yet the capabilities of LLMs have not been integrated into commenting features of document reader software. Instead, most LLM-enhanced document readers require users to summarize and ask questions about the text in a separate, chat-like interface disconnected from the document. There is an opportunity to improve how LLMs are used in this context by integrating their capabilities directly into ‘margin note’ comments that are linked to the document text.

	Where	When
R	CH. 3	CH. 4
W		CH. 5
R+W	CH. 6	

*This project explores constraining **where** prompting can occur and **when** comments are finalized while **reading and writing** in a document.*

We have already seen a lot of margin notes in this dissertation!

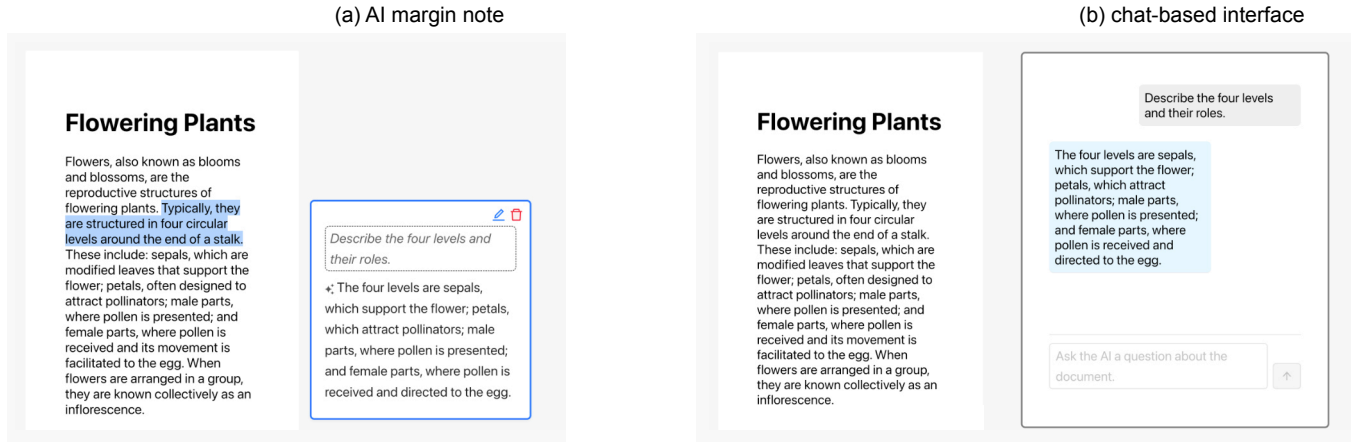


Figure 6.1: Different ways to leverage LLM capabilities in document reader software: (a) we propose AI margin notes that are integrated with the document text unlike (b) chat-based interfaces that are separated from the document text.

We explore the design of “AI margin notes” (Figure 6.1) that leverage LLMs to enhance comments in document reader software through three controlled experiments. Every experiment required participants to read short, non-fiction documents while interacting with an LLM, primarily within comments, and answer reading comprehension questions two hours later. Each experiment focused on a different design parameter. First, we compared AI margin notes to traditional, chat-based prompting to better understand the effects of *integration*. Second, we compared manually selecting text that an AI margin note is associated with, to automatically selecting text through an LLM-powered assistant to better understand the effects of *selection automation*. Third, we explored different AI margin note techniques that leverage LLMs in different ways to understand the effects of *human and AI involvement*. Specifically, AI margin notes that generate summaries, fill-in-the-blank exercises, responses to specific prompts, practice short answer questions, and feedback on how written text can be improved.

Our results indicated that AI margin notes were preferred over chat interfaces and that selecting text for AI margin notes should be done manually. AI margin note techniques with varying levels of human and AI involvement were valued for different reasons. For example, techniques with more human involvement were typically associated with more *psychological ownership*, but techniques with more AI involvement to finalize the note were faster, less effortful, and generally preferred. Our work contributes the idea of AI margin notes and empirical results demonstrating that AI margin notes are a desirable feature for document reader software.

6.2 BACKGROUND AND RELATED WORK

AI margin notes relate to existing literature in psychology about the benefits and challenges of note-taking while reading, and other techniques that have used LLMs to improve reading and note-taking. We focus specifically on

relevant work focused on reading, rather than taking notes while attending a lecture or watching a video.

6.2.1 *Note-Taking while Reading*

There are many types of notes that readers can create, margin notes being one of them. Research in psychology suggests that the general activity of note-taking is beneficial for two main reasons [91]. First, notes act as external storage for information, which can be re-read to reinforce memory (the *storage function*). Second, the act of creating notes can facilitate learning as it requires readers to pay more attention to the material to process and organize ideas (the *encoding function*).

The encoding function is especially beneficial when the reader processes the text at a deeper level [47, 90], for example, by connecting it to prior knowledge and experiences (*levels of processing theory*). However, many people opt for shallower and less effective note-taking strategies. For example, Bretzing and Kulhavy’s analysis of students’ note-taking activities [28] revealed that students tend to take verbatim notes that repeats text from documents. In another study [27], they showed that students who took verbatim notes performed worse on a test than those who used deeper note-taking strategies by writing their own summaries or by paraphrasing text from the document.

Note-taking is a complex activity that requires significant cognitive effort as readers must coordinate and frequently switch between reading and writing activities. Writing takes more time than reading, and excessive delays between reading activities as a result of note-taking can hinder comprehension. Therefore, readers often experience significant mental and temporal demand, even when they are reading and taking notes without any time limits [140]. This can become especially tiring with longer [96] or poorly-formatted documents [132]. Although researchers have argued that increased cognitive effort can benefit learning [20–22] and memory [173], it does not always lead to improved learning outcomes, for example, if the task is too frustrating or if the learner lacks motivation [67].

6.2.2 *Reading and Note-Taking with LLMs*

Recent work has investigated how LLMs can improve user experience and comprehension while reading and taking notes. For example, text simplification techniques that turn complex documents into simplified versions [69] and make documents easier to skim [68] can improve reading comprehension and reduce workload. Users of systems like ChatGPT, Adobe Acrobat [3], Google Notebook LM [64], NoteGPT [130], and ChatPDF [41] can ask questions, summarize, and write notes about documents. However, few investigations examine the effect on factors like reading comprehension, reading duration, workload, and preferences.

Kreijkes et al. [98] asked high school students to try different note-taking techniques to study for reading comprehension tests that took place three days later: note-taking independently and note-taking while having access to an LLM-powered chatbot to ask questions. Both of these note-taking

techniques were compared to a baseline of asking the chatbot questions about the document, without any note-taking. Their results showed that both note-taking techniques were associated with better performance than just asking the chatbot questions about the document, suggesting that using LLMs in a more cognitively engaging way (i.e., with some note-taking) can improve reading comprehension. However, they did not compare the two note-taking conditions, making the effect of LLM use on note-taking unclear.

Although Kreijkes et al. found that just asking the LLM questions resulted in poorer test performance, it was preferred by participants as it was perceived to be more enjoyable and less effortful. Such findings have been reproduced in other studies focused on the effects of LLMs on learning more broadly. Systems like ChatGPT can allow for personalized learning experiences, which can improve academic performance, reduce mental effort, and motivate learners [50, 177], however, learners may over-rely on these systems [177]. A balance may be to encourage more cognitive engagement. For example, Chat-PRCS [179] generates practice reading comprehension questions for students, which increased mental load but improved reading comprehension. Similarly, CoAsker [110] encourages students to generate practice questions with an LLM-powered assistant, which were displayed in a side panel much like notes beside the document. When compared to generating practice questions without assistance, receiving questions that were generated by the assistant increased reading comprehension scores.

To our knowledge, no prior work has thoroughly explored integrating LLMs into the commenting feature of document reader software. The closest prior work are prototype mockups in Melin-Higgins' bachelor's thesis [119], in which users select document text to issue pre-determined prompts to a hypothetical LLM-powered assistant. Responses appear as 'sticky notes' beside the selected text, like margin notes, or they appear underneath the selected text by modifying the document structure. A very small 3-person study showed a preference for the margin note style. The study also only focused on usability and user preferences, and did not consider factors like reading comprehension. Furthermore, the prototypes were partially functioning Figma mockups, and none explored different levels of human and AI involvement.

Based on this research, AI margin notes could hinder or help readers. Creating them may limit deeper processing of the document text and lower comprehension. However, some AI margin note creation techniques may make note-taking less mentally demanding, so readers focus more on the underlying text while reading. Some techniques may even help readers take non-verbatim notes while improving motivation and cognitive engagement. Therefore, it is important to understand the effect of AI margin notes, and note-taking with LLMs more generally, on factors like test performance, workload, and user preferences. Our work contributes these important insights, which have been lacking in existing literature.

6.3 AI MARGIN NOTES

We focus on three design parameters of AI margin notes: *integration*, *selection automation*, and the level of *human and AI involvement*.

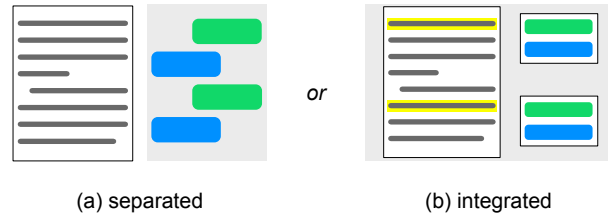


Figure 6.2: Integration: (a) a chat-based interface is not integrated since it is separated from the associated document text and (b) AI margin notes are integrated since they are associated to specific document text. *In this and following two figures: green denotes human-written text, like a prompt, and blue represents AI-generated text, like a response. Yellow indicates text that the AI margin note is linked to.*

6.3.1 Integration

The primary difference between AI margin notes and chat-based tools is integration. Specifically, AI margin notes enhance *comments* in document readers. Comments are anchored to specific text, making them integrated into the document, which contrasts with chat-based tools that are placed in a separate side panel with the content disconnected from the context (Figure 6.2). A separate interface for interacting with an AI assistant may suffice for general questions about the document or to receive overall summaries. However, readers also direct the AI assistant to specific parts of the text to contextualize their prompt [98]. Separating these responses from the text can be inefficient.

First, *referring to specific parts of a document while prompting requires additional work* [118]. For example, if the user wants to simplify the language of a paragraph in a scientific document (e.g., [69]), they must first type a command like “simplify this,” select the paragraph from the document, then copy and paste it after the command. This copy and paste step could be avoided entirely if the user selected the paragraph and typed “simplify this” in an AI margin note. Or easier, by just selecting text and pressing a button that reifies the “simplify this” prompt [9, 118].

Second, *shifting attention to a separate interface may distract from reading*. Note-taking requires frequent switching between reading and writing [3, 140], and researchers suggest that note-taking should “interrupt reading as little as possible” [115]. Having readers frequently switch between separate reading and writing interfaces may further increase cognitive load [73]. These switching costs could be mitigated by presenting the prompting interface alongside the specific text the user is reading.

Third, *referring back to specific responses requires additional work*. For example, suppose a user used a chat-based LLM interface to get an explanation for an unfamiliar concept in a document. The explanation may be remembered in the short term, but if the user revisits the document much later, they must scroll

through a lengthy and poorly organized chat history to find it, which may be especially difficult for readers with a lower [working memory capacity](#) to do [154]. Linking responses directly to the relevant text avoids this issue. Some systems, like Adobe Acrobat Reader, support clicking on chat responses to highlight relevant parts of the document, however, this still requires scrolling through chat history. Alternatively, the user could re-issue the prompt, but this wastes time and resources. Of course, users could copy responses they wish to save and paste them in comments [98], but this requires additional copy and paste steps that would not be needed with an AI margin note.

6.3.2 Selection Automation

AI margin notes are associated with specific text, which defines the context for the prompt and a location to display the note. Specifying text for the note could be done automatically by the AI assistant or manually by the user (Figure 6.3). For example, consider summarization, a common task readers do in conventional margin notes [80, 143] and with LLM systems [98, 119]. When using LLMs in document reader software, readers can generate a summary of the entire document automatically by pressing a single button [3], or they can manually specify which parts of the document need to be summarized by copying and pasting text into the chat. When the reader just presses a button, the LLM decides which information is important and worth including in the summary, but when the reader copies text to summarize, they are deciding which parts of the document are most important. Prior work on text highlighting suggests that this decision process could improve learning outcomes and reading comprehension, but it could also increase mental effort [192].

Automating the selection of text to create AI margin notes may also impact [psychological ownership](#), which we discussed in [Chapter 5](#). Prior work shows that learners can feel more psychological ownership if they are given more *control* over their learning [138, 139], which can encourage more active learning [65] and increase motivation [165]. Automatically generating AI margin notes reduces effort, but potentially at the risk of lowering comprehension and psychological ownership.

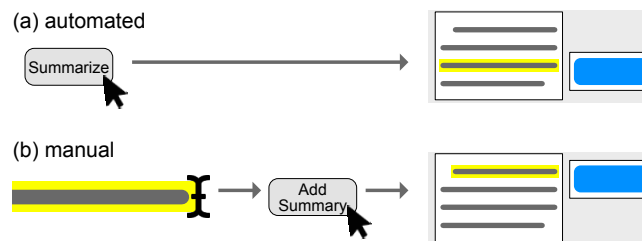


Figure 6.3: Selection automation: (a) text can be automatically selected and multiple AI margin notes created at once and (b) the user can manually select text to create an AI margin note.

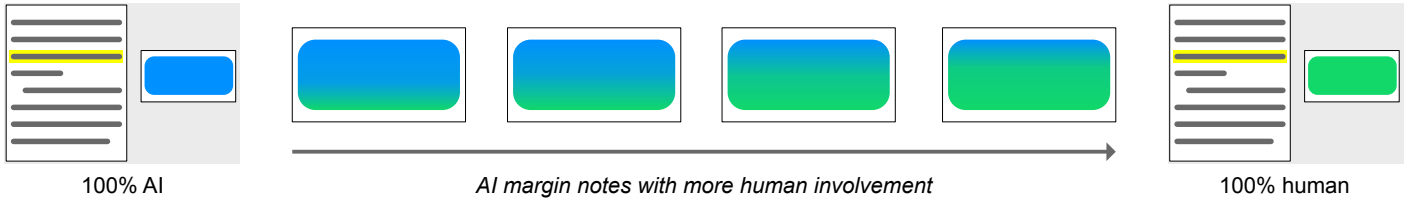


Figure 6.4: Levels of human and AI involvement: traditional margin notes consist of text that is 100% written by a human, but an AI margin note could consist of text that was 100% generated by AI, or could involve more human-generated text.

6.3.3 Human and AI Involvement

Currently, commenting in document reader software is intended for text that is 100% written by a human. Specifically, the user must formulate their own ideas and type into a text box. At the other extreme, pressing a “Summarize” button in a document reader [3] produces text that is 100% written by the LLM and does not require any guidance from the human. However, there are techniques for producing text that require involvement from both the human and the LLM. For example, chat-based interfaces require the user to write a prompt so the LLM can produce relevant output text. Here, the user exerts control over what the output should be, and the LLM produces it (Chapter 5). These roles can be reversed too, for example, the LLM could ‘prompt’ the user to respond to practice reading comprehension questions [110, 179].

Based on these existing variations of human and AI involvement when interacting with LLMs, we believe that AI margin notes could also vary in how much human and AI involvement they require to produce the final comment text (Figure 6.4). Techniques that require more human involvement likely require more cognitive engagement than those with more AI involvement, but this may be beneficial for reading comprehension [20–22, 173], may discourage readers from taking verbatim notes [27], and may even improve feelings of psychological ownership. However, frustrating cognitive engagement can discourage learners [67], which may be mitigated by increasing AI involvement.

6.4 EXPERIMENTAL METHOD

We conducted a series of experiments to explore the design parameters of AI margin notes. Specifically, we were interested in understanding how integration, selection automation, and human and AI involvement affect reading comprehension and duration, psychological ownership, task workload, and user preferences. All experiments were conducted using Prolific.

6.4.1 Task

The primary experimental task was composed of two stages. First, a *reading stage*, where participants read non-fiction documents that were approximately 500 words each. For each document, they interacted with an LLM assistant

using a different technique that represents a variation within the design parameter under evaluation. Second, a *test stage*, where participants completed reading comprehension tests about each document. Each test consisted of six multiple choice questions. The documents and questions were developed by Wallace et al. [175] and were designed by a learning and reading specialist to be suitable for an 8th grade reading level.¹ This is representative of documents targeted for the general public [126].

6.4.2 Apparatus

Our experimental system was a custom Node.js and React web application that implemented two types of interfaces, one for each stage: the reading interface, where participants read a document and interacted with an LLM using a specific technique (Figure 6.5); and the testing interface, where participants answered time-bounded, multiple-choice reading comprehension tests.

6.4.2.1 Reading Interface

The document was displayed on the left side of the screen. The specific design variation of the experimental condition was displayed beside it (details provided in each individual experiment). All design variations that involved an LLM used GPT 4.1 mini. The top of the screen contained a toolbar indicating how many comments or responses they still had to complete and a blue “Finish Reading” button to end the trial.

6.4.2.2 Testing Interface

The testing interface displayed six multiple choice questions at the centre of the screen. At the top of the screen was a toolbar that displayed how many questions the participant had answered, a countdown timer (displayed numerically and as a progress bar that shrank every second), and a blue “Finish Test” button that could be pressed to end the test.

6.4.3 Procedure

Participants received a link to an experimental web application through Prolific. They had to use a desktop or laptop computer, which was strictly enforced through the web application. First, they entered basic demographic information and read instructions. Next, they completed the reading stage for all documents. There was no time limit for the reading stage, however, to focus on the effect of the different techniques rather than the number of responses, participants could only leave 3 AI margin notes (or receive 3 responses from the AI assistant, depending on the technique). After reading the document, they answered 10 questions about their experience and 1 question about

¹ Note that we slightly modified the wording of some questions, as we noticed that some questions could be successfully answered by looking at the wording of other questions. At the time of writing this dissertation, Wallace et al.’s license agreement permits reuse, modification, public display, and redistribution for non-commercial research purposes.

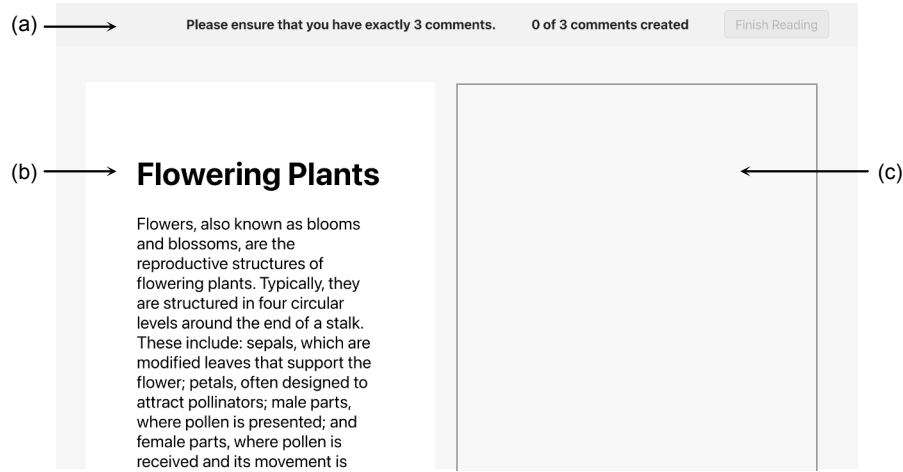


Figure 6.5: Experimental reading interface: (a) toolbar containing instructions, (b) document to read, and (c) space for specific design variations to be displayed.

their prior knowledge of the document’s content. They repeated this for the other documents, then answered 2 questions about their overall preferences. They also described other study aids they used for the test, such as taking a screenshot of the document or writing notes outside of the reading interface.

Two hours later, the participant was invited to return for the test stage. Given the lower reading difficulty of the documents, we increased the difficulty of the test by making it closed-book and restricted to 60 seconds.

6.4.4 Design

All experiments use within-subjects design since the documents were relatively short, and we wanted participants to compare their experiences with each technique. There is one independent variable, *CONDITION*, which was randomly assigned. For a single *CONDITION*, one of six documents from Wallace et al. [175] was randomly-assigned. For increased internal validity, a single document was not restricted to a single condition and could be assigned to any condition.

6.4.5 Measures

Reading comprehension and duration were calculated from logs, and all other metrics were calculated from subjective questionnaires that participants completed after the reading stages. All used a 0-100 interval range.

6.4.5.1 Reading Comprehension

This represents the number of questions that were correctly answered during the test stage (0-6 range). Analyzing *Reading Comprehension* gives insights into the effects of AI margin notes on overall learning, and whether techniques that encourage more human involvement improve learning outcomes.

6.4.5.2 *Duration*

This is the time taken in minutes to complete the reading stage. Some AI margin notes that require more manual effort or human involvement likely take longer to complete.

6.4.5.3 *Psychological Ownership*

This was measured by asking two quantitative questions about *Personal Ownership* and *Responsibility* [36]. The average of the two create a composite measure, as done in Chapter 5. The internal consistency reliability score was high for all experiments ($.80 \leq \alpha \leq .90$), suggesting that the composite measure was appropriate to use for data analysis. As described previously, this is an important measure since fostering psychological ownership can improve learner experiences.

6.4.5.4 *Task Workload*

We used factors from the NASA-TLX: *Mental Demand*, *Physical Demand*, *Temporal Demand*, *Performance*, *Effort*, and *Frustration*. Techniques that are more manual or that require more human involvement may involve more mental demand and effort, which could be beneficial for learning [20–22, 173], however, this may not happen if they are too frustrating [67].

6.4.5.5 *Preferences*

We asked about *Frequency of Use*, a question from the system usability scale (SUS) representing how frequently the participant would use the feature if made available to them. After trying all techniques, participants also gave each CONDITION a *Ranking*, where 1 was the best, 2 was the second best, and so on. Ties were allowed. To establish an *Overall Ranking* of conditions, we use the Condorcet voting method [190]. An overall rank of 1 means that technique ‘defeats’ all others in pairwise comparisons. An overall rank of 2 means that technique ‘defeats’ all others except for the technique ranked 1, and so on. This is important to study, as users would likely want to keep using techniques that they like, which is especially useful if it also improves their learning outcomes.

6.5 EXPERIMENT 1: INTEGRATION

The goal of this experiment is to understand the effect of integration that is achieved with AI margin notes. Participants wrote prompts using a traditional, chat-based interface (CHAT), or an AI margin note (NOTE).

6.5.1 *Participants*

We recruited 29 participants through Prolific. Participants were restricted to the United States and Canada, and those who had completed 2,500 previous tasks on the platform with a 99-100% approval rating. Three participants

Table 6.1: Demographics for Experiment 1.

Gender		Age		Education	
Men	16	25-34	4	High School	2
Women	10	35-44	14	Some University (no credit)	5
		45-54	3	Bachelor's Degree	14
		55-64	3	Professional Degree Beyond Bachelor's	1
		65-74	2	Master's Degree	3
				Doctorate Degree	1

Document Reader Frequency		Commenting Frequency		LLM Frequency		LLM Summarization Frequency	
Daily	6	Daily	1	Daily	10	Daily	3
Weekly	11	Weekly	6	Weekly	9	Weekly	9
Monthly	7	Monthly	3	Monthly	5	Monthly	3
Less than Monthly	2	Less than Monthly	5	Less than Monthly	5	Less than Monthly	5
		Never	11	Never	1	Never	6

(10%) were removed for using other study aids prior to the test, leaving 26 valid responses (Table 6.1). All self-reported proficiency in reading in English. Participants received \$10 in total, and the top 25% scoring participants received an additional \$5 bonus to incentivize them to do well on the test. The experiment took roughly 25 minutes in total.

6.5.2 Apparatus

The reading interface had two variations. For the chat-based prompting technique, a chat interface was displayed to the right of the document. Here, participants could type prompts in a text box and press a blue “Submit” button. Their prompts appeared like a chat message on the right in a grey bubble, and LLM responses were displayed on the left in a light blue bubble (Figure 6.6a).

To make more direct comparisons between the two variations, we had to create an AI margin note technique where participants type prompts and receive responses from an LLM. To create an AI margin note, the participant selected text in the document using their cursor, which caused a blue “Comment” button to appear. Clicking this created a text box to appear that was anchored to part of the document text and resembled a Google Docs comment (Figure 6.6b). The participant could type a prompt and press a blue “Confirm” button. After a few seconds, the LLM provided a response. The participant’s prompt appeared at the top of the comment in grey, italicized text with a dashed border around it, and the response appeared below in black text, prepended with a black ‘sparkle’ icon. This styling clearly distinguishes human-written text (no icon) from AI-written text (with a ‘sparkle’ icon) and main comment text (black text) from additional context that helped generate the comment (grey, italicized text with a dashed border). Participants could edit their prompt to receive new output and they could delete their comments. Clicking on individual comments highlighted the corresponding text in the document.

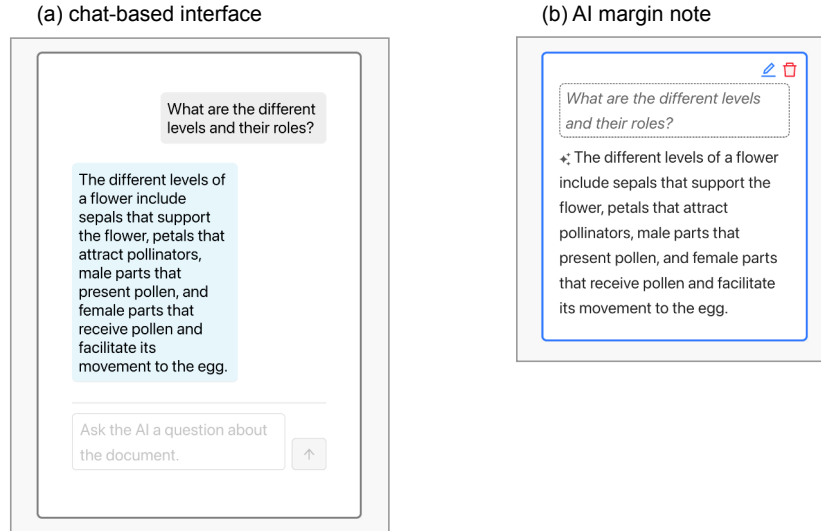


Figure 6.6: Techniques tested in Experiment 1: (a) a chat-based interface and (b) AI margin notes.

6.5.3 Results

We use Wilcoxon signed-rank tests to investigate the effects of *CONDITION* on the various measures as the data was non-parametric. Details of these statistical tests are shown in [Table 6.2](#).²

Overall, there were no significant effects of *CONDITION* on *Reading Comprehension*, *Duration*, *Psychological Ownership*, any of the workload-related factors, and *Frequency of Use*.

Table 6.2: Statistical test results for Experiment 1.

Measure	<i>W</i>	<i>p</i>	<i>RBC</i>
<i>Reading Comprehension</i>	80	.81	.06
<i>Duration</i>	148	.50	.16
<i>Psychological Ownership</i>	82.5	.40	.21
<i>Mental Demand</i>	149	.72	-.08
<i>Physical Demand</i>	76	.68	.11
<i>Temporal Demand</i>	74.5	.63	-.13
<i>Performance</i>	70.5	.78	.08
<i>Effort</i>	105.5	.32	.24
<i>Frustration</i>	57	.12	-.40
<i>Frequency of Use</i>	77.5	.07	.44

² Note that *W* here refers to the Wilcoxon signed-rank test statistic, not Kendall's *W* like we saw in [Chapter 5](#).

6.5.3.1 Preferences

Although there were no differences for the aforementioned measures, most participants preferred prompting with AI margin notes. From the Condorcet voting method, we observed that NOTE received an *Overall Ranking* of 1, and CHAT received an *Overall Ranking* of 2. Notably, majority of participants (17, 65%) assigned NOTE a *Ranking* of 1, and CHAT a *Ranking* of 2 (16, 62%).

To better understand why participants preferred NOTE over CHAT, we examined participants’ free-form responses. Several (12, 46%) mentioned how NOTE was easier and more intuitive than CHAT. Participants described how they “liked [not] having to move to a separate chat” (P21), how “selecting text [was] more natural when one has a question related to that part of the text” (P14), and how “having the comment embedded right into the document [was] handy and easy to reference along with the actual text” (P5).

Overall, our results suggested that the integration of AI margin notes is highly desirable and preferable over chat-based interfaces as it prevented switching between interfaces, it was easier for readers to indicate text they had questions about, and it was to reference individual comments later. In this experiment, associating text to the AI margin note was done manually by the participant. However, selecting text could also be done automatically by the LLM, which we explore further in the following experiment.

6.6 EXPERIMENT 2: SELECTION AUTOMATION

The goal of this experiment is to understand the effect of selection automation when creating AI margin notes. Participants either manually selected what text should be summarized (MANUAL), or pressed a button to let the AI assistant place three summary AI margin notes in the document (AUTOMATIC).

6.6.1 Participants

Using the same inclusion criteria as Experiment 1, we recruited 32 new participants on Prolific. Two participants (6%) were excluded for not attempting to answer any reading comprehension questions, or for using other study aids, leaving 30 valid responses (Table 6.3). All self-reported English reading proficiency. Like before, participants received \$10 as remuneration, with the top 25% receiving an additional \$5 bonus. The experiment took approximately 25 minutes in total.

6.6.2 Apparatus

Both variations focused specifically on producing AI margin notes that summarized text as this is a representative task that is currently done manually, by the user copying and pasting specific parts of a document into a chat-based interface, or automatically, by pressing a button. For the automatic selection technique, the top toolbar displayed a blue “Generate Comments” button in the top left corner (Figure 6.7a). Pressing this caused three AI margin note comments that were linked to the document text to appear to the right of the

Table 6.3: Demographics for Experiment 2.

Gender		Age		Education	
Men	18	25-34	4	High School	4
Women	12	35-44	15	Some University (no credit)	7
		45-54	8	Bachelor's Degree	9
		55-64	1	Master's Degree	10
		65-74	1		
		75+	1		

Document Reader Frequency		Commenting Frequency		LLM Frequency		LLM Summarization Frequency	
Daily	4	Daily	1	Daily	8	Daily	2
Weekly	10	Weekly	3	Weekly	12	Weekly	8
Monthly	7	Monthly	4	Monthly	5	Monthly	4
Less than Monthly	7	Less than Monthly	3	Less than Monthly	5	Less than Monthly	11
Never	2	Never	18	Never	2	Never	5

document. Each summarized specific selections from the document, and was displayed in black text prepended with a black ‘sparkle’ icon. As before, participants could click individual comments to highlight the corresponding text, but they could not delete or edit individual comments. Instead, they could press the “Generate Comment” button again to regenerate all comments.

The manual selection technique worked like the previous experiment (Figure 6.7b): the participant selected text from the document and pressed a blue “Comment” button to generate a summary comment (appearing after a few seconds). The styling of the comment was the same, except that participants could also delete or regenerate individual comments. The participant had to manually create three summary comments.

6.6.3 Results

As before, we use Wilcoxon signed-rank tests where applicable. Details of statistical tests are shown in Table 6.4.

We did not observe any significant effect of CONDITION on *Reading Comprehension*, so we focus our results on other metrics.

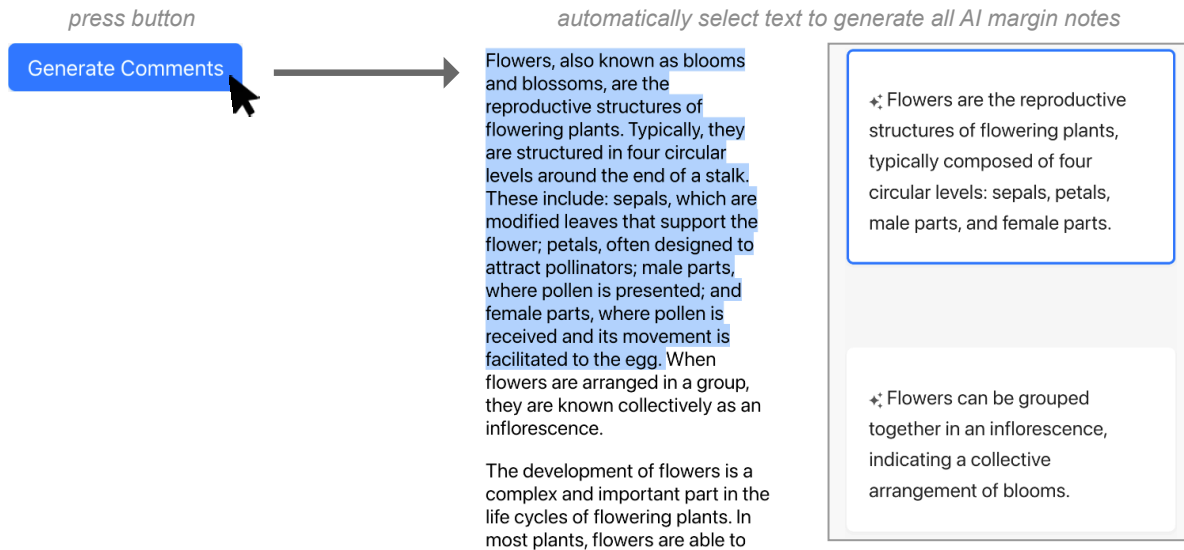
6.6.3.1 Duration

Participants were 39 seconds slower when they manually selected text (Figure 6.8a). There was a significant effect of CONDITION on *Duration*, revealing that MANUAL (MDN = 3.43, IQR = 2.15) was slower than AUTOMATIC (MDN = 2.78, IQR = 2.14).

6.6.3.2 Psychological Ownership

Though slower, participants generally felt more psychological ownership when they manually selected text (Figure 6.8b). A significant effect of CONDITION on *Psychological Ownership* revealed that participants felt more *Psycho-*

(a) automatic text selection



(b) manual text selection

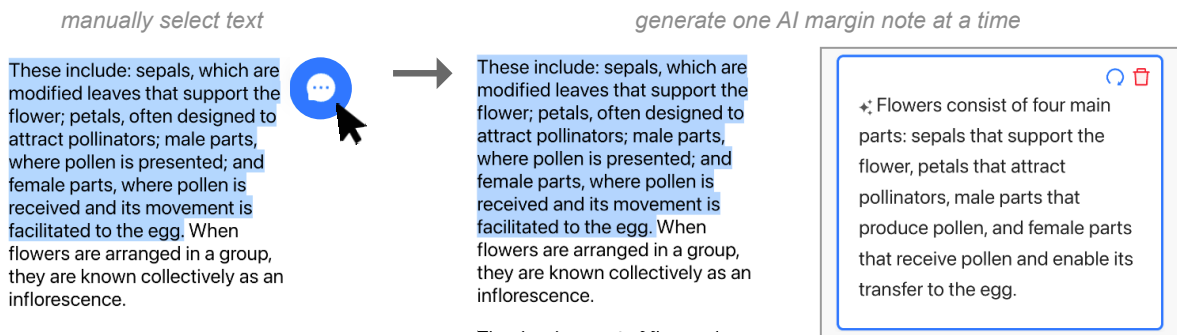


Figure 6.7: Techniques tested in Experiment 2: (a) automatically selecting text to generate all AI margin notes at once and (b) manually selecting text and generating AI margin notes one-by-one.

logical Ownership with MANUAL (MDN = 50.25, IQR = 46.62) than AUTOMATIC (MDN = 6.75, IQR = 10.88). This was supported by free-form responses like “[manually selecting text] requires me to focus more on the task, [which] gives me more responsibility” (P20).

6.6.3.3 Task Workload

Participants felt like they performed better when selecting text manually, despite it requiring more effort. Specifically, there was a significant effect of CONDITION on *Performance* (Figure 6.8c), suggesting that participants felt like they were better at creating AI margin notes with MANUAL (MDN = 88, IQR = 24.25) than AUTOMATIC (MDN = 80.5, IQR = 49.5).

A significant effect of CONDITION on *Effort* (Figure 6.8d) suggested that participants felt more *Effort* for MANUAL (MDN = 35, IQR = 41.25) than AUTOMATIC (MDN = 14, IQR = 38.75). This was supported by free-form responses

Table 6.4: Statistical test results for Experiment 2.

Measure	W	p-value	RBC
Reading Comprehension	88.5	.20	.30
Duration	102	.006 **	.56
Psychological Ownership	9	< .001 ***	.95
Mental Demand	158.5	.46	.16
Physical Demand	31.5	.10	.48
Temporal Demand	100	.25	.28
Performance	52.5	.02 *	.58
Effort	84.5	.004 **	.61
Frustration	77.5	.19	.33
Frequency of Use	63.5	.004 **	.64

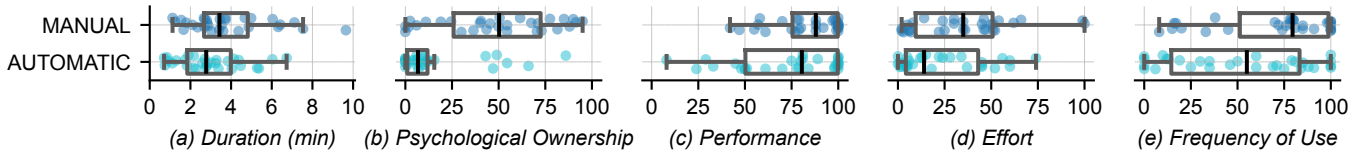


Figure 6.8: Experiment 2 results: (a) Duration, (b) Psychological Ownership, (c) Performance, (d) Effort, and (e) Frequency of Use.

like “in being able to highlight on my own, I also feel as if I did some of the work and not just nothing” (P25). We did not observe any differences between MANUAL and AUTOMATIC for *Mental Demand*, *Physical Demand*, *Temporal Demand*, and *Frustration*.

6.6.3.4 Preferences

Participants generally seemed to prefer generating AI margin notes manually. There was a significant effect of CONDITION on *Frequency of Use* (Figure 6.8e), suggesting a preference to use MANUAL (MDN = 79.5, IQR = 47.5) more frequently than AUTOMATIC (MDN = 55, IQR = 68.75). For *Overall Ranking*, we observed that participants tended to prefer MANUAL, which was ranked first. Specifically, 23 participants (77%) gave MANUAL a *Ranking* of 1, and 21 (70%) gave AUTOMATIC a *Ranking* of 2.

Majority (16, 53%) appreciated MANUAL for the increased control it enabled, for example: “I felt a little more control [when] I was able to pick what I wanted to be summarized. [When] I just pressed one button, it summarized everything, but not necessarily the areas I wanted it to” (P22). A few (4, 13%) mentioned how MANUAL provided more incentive to read the document, for example: “generating all summaries at once will encourage users not to actually read the document for themselves. Generating individual summaries after I selected text meant I [had] to actually read the document” (P10).

Overall, our results suggested that selecting text for an AI margin note is slower and requires more effort when done manually. However, this may be a worthwhile trade-off, as manually selecting text was associated with higher

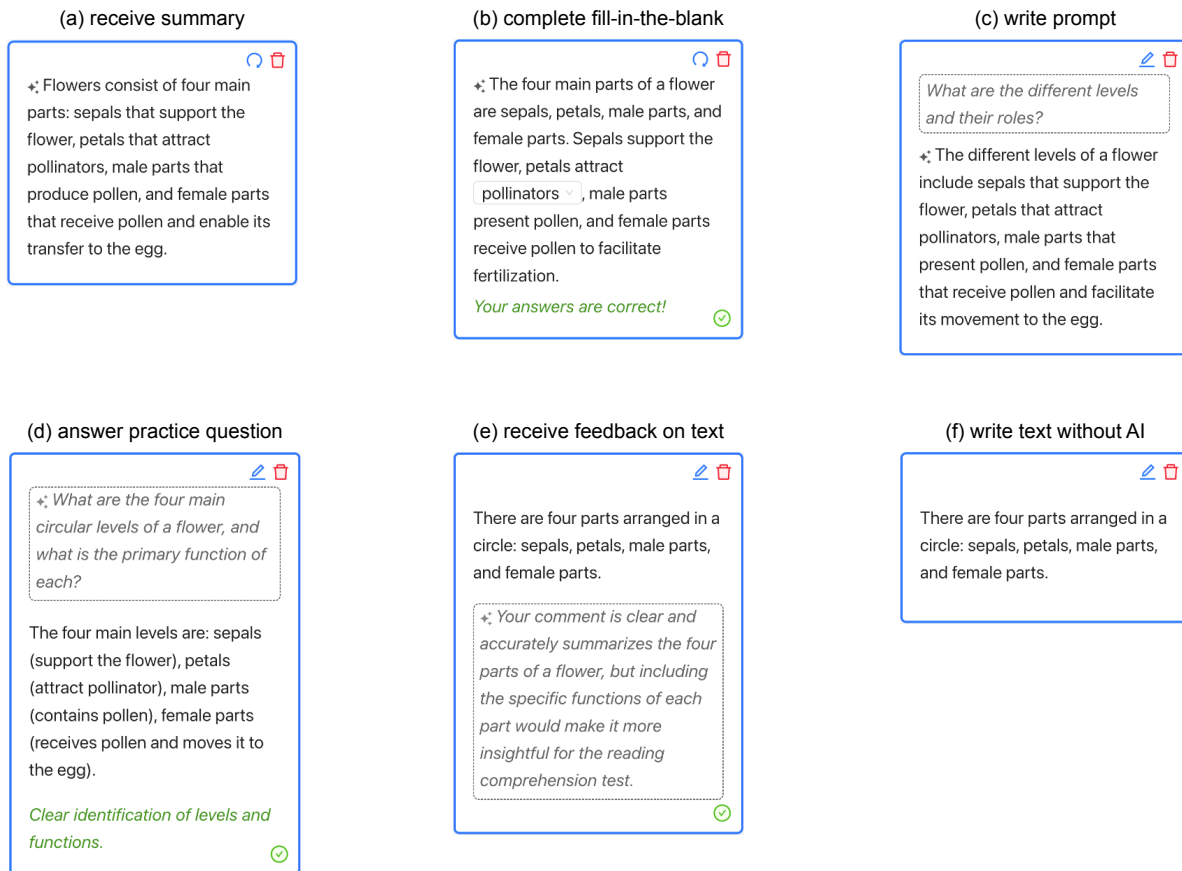


Figure 6.9: Techniques tested in Experiment 3: (a) receive a summary (100% AI-written text), (b) complete a fill-in-the-blank exercise, (c) write a prompt, (d) answer a practice question, (e) receive feedback on written text, and (f) write a comment without AI (100% human-written text).

feelings of psychological ownership, better perceived performance, and was preferred by participants. In this experiment, the resulting AI margin note was always a summary, however, there are many other types of AI margin notes that can be created that involve the human and AI in different ways.

6.7 AI MARGIN NOTE TECHNIQUES

To fully explore the design parameter of human and AI involvement, we implemented six techniques that fall along a continuum, ranging from AI margin notes whose text was generated entirely by the LLM, to those whose text was written entirely by the user (Figure 6.9). All are created by selecting text manually and pressing a blue “Comment” button.

6.7.1 *Receive Summary*

The technique with the least human involvement is receiving a generated summary where 100% of the text is written by AI (Figure 6.9a). This is identical to the manual text selection technique presented in Experiment 2.

6.7.2 *Complete Fill-in-the-Blank Exercise*

A technique with slightly more human involvement is receiving a generated summary, but with some keywords omitted to resemble a fill-in-the-blank question (Figure 6.9b). Like the previous technique, the LLM produces a summary for the text selected by the user, shown in black text and prepended by a black ‘sparkle’ icon. Each summary contains one or two ‘blanks,’ displayed using dropdowns containing three options. The user must select the correct answer from the provided options. Underneath, the comment displays feedback to notify the user whether their responses are correct (shown in green with a check mark) or incorrect (shown in red with an ‘X’). The user can delete the comment or regenerate it.

6.7.3 *Write Prompt*

A technique with some human and some AI involvement is writing a custom prompt for the LLM to respond to (Figure 6.9c). Writing a prompt allows the user to exert more control over the output, for example, to achieve goals that go beyond summarization such as learning additional information about the text or simplifying it. This is identical to the AI margin note technique described in Experiment 1.

6.7.4 *Answer Practice Question*

Another technique with some involvement from both the human and the AI is answering a practice reading comprehension question that was generated by the LLM, which has been shown to help with learning [110, 179]. The LLM produces a practice short-answer question based on the selected text (Figure 6.9d), prepended with a grey ‘sparkle’ icon and shown using grey italicized text with a dashed border. The user types in their answer in a text box underneath and presses a blue “Confirm” button to submit their response. The submitted response is then shown in black text. After a few seconds, the user receives feedback from the LLM about their answer: correct responses are shown in green and with a check mark underneath the response, and incorrect responses are shown in red with an ‘X.’ All responses have to be correct. The user can edit their responses to correct mistakes (and will receive new feedback accordingly), or can delete their comment.

6.7.5 *Receive Feedback on Written Text*

A technique that has a lot of human involvement is requiring users to write their own comment text, with LLM-generated feedback and suggestions on

how their comment can be improved (Figure 6.9e). This can be beneficial, as many learners struggle to take effective notes [27]. The user can type within the provided text box and press “Confirm.” The comment text is shown as black text. After a few seconds, the AI assistant provides feedback on the text, prepended with a grey ‘sparkle’ icon and shown as grey, italicized text with a dashed border. A green check mark is shown if their comment text is ‘good,’ and a red ‘X’ is shown if it is not. All comment text has to be classified as good. The user can edit their comment to receive updated feedback and can delete it.

6.7.6 Write Text without AI

The technique with the most human involvement is requiring the user to write their own comment, without any AI assistance (Figure 6.9f). This baseline resembles the behaviour of existing commenting systems, where 100% of the text is written by the user. The user can type within the provided text box and press “Confirm.” The comment is shown as black text. The user can edit or delete the comment.

6.8 EXPERIMENT 3: HUMAN AND AI INVOLVEMENT

The goal of this experiment is to investigate the effect of human and AI involvement. Participants read six documents, one for each AI margin note technique described above: receiving a summary (SUMMARY), completing a fill-in-the-blank exercise (BLANK), writing a prompt (PROMPT), answering a practice question (QUESTION), receiving feedback on written text (FEEDBACK), and writing without any AI assistance (NONE). For BLANK, QUESTION, and FEEDBACK, their answers and responses had to be marked as ‘correct’ or ‘good,’ restricting *when* they could continue. The experiment was longer, approximately 90 minutes.

Table 6.5: Demographics for Experiment 3.

Gender		Age		Education	
Men	13	18-24	1	High School	7
Women	19	25-34	2	Some University (no credit)	2
		35-44	7	Technical Degree	5
		45-54	13	Bachelor’s Degree	14
		55-64	4	Master’s Degree	4
		65-74	4		
		75+	1		

Document Reader Frequency		Commenting Frequency		LLM Frequency		LLM Summarization Frequency	
Daily	6	Daily	1	Daily	13	Daily	3
Weekly	14	Weekly	6	Weekly	11	Weekly	12
Monthly	8	Monthly	4	Monthly	5	Monthly	3
Less than Monthly	3	Less than Monthly	8	Less than Monthly	3	Less than Monthly	10
Never	1	Never	12			Never	4

6.8.1 Participants

We recruited 36 new participants on Prolific, using the same inclusion criteria as the previous experiments. Four (11%) were excluded for experiencing technical issues, leaving 32 valid responses (Table 6.5). All but one were proficient at reading in English. Participants received \$25 total, with an additional \$5 bonus if their total reading comprehension score was within the top 25%.

6.8.2 Results

Where applicable, we use Friedman omnibus tests and Wilcoxon signed-rank post hoc tests, with Holm’s corrections for multiple comparisons. Statistical test details are shown in Table 6.6 and Table 6.7.

Table 6.6: Omnibus statistical test results for Experiment 3.

Measure	χ^2	<i>p-value</i>		Kendall’s <i>W</i>
<i>Reading Comprehension</i>	10.87	.05		.07
<i>Duration</i>	64.64	< .001	***	.40
<i>Psychological Ownership</i>	74.76	< .001	***	.47
<i>Mental Demand</i>	57.83	< .001	***	.36
<i>Physical Demand</i>	20.92	< .001	***	.13 <i>post hocs were n.s.</i>
<i>Temporal Demand</i>	4.45	.49		.03
<i>Performance</i>	16.82	.005	**	.10
<i>Effort</i>	55.98	< .001	***	.35
<i>Frustration</i>	5.39	.37		.03
<i>Frequency of Use</i>	9.19	.10		.06

6.8.2.1 Reading Comprehension

Contrary to what prior work suggests, techniques with more human involvement were not associated with higher reading comprehension scores. Specifically, the differences between SUMMARY (MDN = 4, IQR = 2.25) and all other conditions (all MDN = 5, IQR = 2) are marginal ($p = .05$), and not statistically significant.

6.8.2.2 Duration

Participants were faster when they used techniques that did not require as much human involvement (Figure 6.10a). A significant effect of CONDITION on *Duration* revealed that SUMMARY (MDN = 3.96, IQR = 2.78) and BLANK (MDN = 4.92, IQR = 2.80) were faster than all other techniques.

Table 6.7: Post hoc statistical test results for significant omnibus tests in Table 6.6: (a) *Duration*, (b) *Psychological Ownership*, (c) *Mental Demand*, (d) *Performance*, (e) *Effort*.

comparisons		(a) <i>Duration</i>		(b) <i>P. Ownership</i>		(c) <i>Mental Demand</i>		(d) <i>Performance</i>		(e) <i>Effort</i>	
		<i>p-value</i>	<i>RBC</i>	<i>p-value</i>	<i>RBC</i>	<i>p-value</i>	<i>RBC</i>	<i>p-value</i>	<i>RBC</i>	<i>p-value</i>	<i>RBC</i>
NONE	FEEDBACK	.02 *	.26	.24	.22	.52	.07	1	.09	.91	.04
NONE	QUESTION	.56	.10	< .001 ***	.54	.06	.27	1	.16	.004 **	.29
NONE	PROMPT	.45	.16	.003 **	.56	.02 *	.33	1	.05	.01 *	.31
NONE	BLANK	.003 **	.47	< .001 ***	.78	< .001 ***	.54	.005 **	.38	< .001 ***	.47
NONE	SUMMARY	< .001 ***	.59	< .001 ***	.89	< .001 ***	.68	1	.17	< .001 ***	.61
FEEDBACK	QUESTION	.18	.19	.002 **	.44	.007 **	.36	1	.07	.01 *	.29
FEEDBACK	PROMPT	.04 *	.38	.003 **	.43	.008 **	.44	1	.04	.01 *	.37
FEEDBACK	BLANK	< .001 ***	.62	< .001 ***	.76	< .001 ***	.60	.03 *	.30	.001 **	.47
FEEDBACK	SUMMARY	< .001 ***	.73	< .001 ***	.85	< .001 ***	.75	1	.10	< .001 ***	.62
QUESTION	PROMPT	.19	.22	.44	.06	.52	.09	1	.11	.91	.05
QUESTION	BLANK	.001 **	.55	.004 **	.46	.05	.32	.08	.23	.03 *	.27
QUESTION	SUMMARY	< .001 ***	.66	< .001 ***	.64	< .001 ***	.53	1	.04	.004 **	.47
PROMPT	BLANK	.03 *	.36	.24	.31	.15	.19	.07	.32	.11	.23
PROMPT	SUMMARY	.002 **	.54	< .001 ***	.54	.02 *	.41	1	.12	.02 *	.40
BLANK	SUMMARY	.14	.27	.003 **	.37	.15	.29	.04 *	.16	.04 *	.28

6.8.2.3 *Psychological Ownership*

Generally, participants felt more psychological ownership for techniques that required more human involvement (Figure 6.10b). A significant effect of CONDITION on *Psychological Ownership* and post hoc tests revealed that NONE (MDN = 100, IQR = 4.38) and FEEDBACK (MDN = 97, IQR = 9.34) were associated with the highest *Psychological Ownership*, and SUMMARY (MDN = 23.75, IQR = 60.88) with the lowest.

6.8.2.4 *Task Workload*

Generally, techniques that required more human involvement were more mentally demanding, associated with poorer perceived performance, and more effortful. For *Mental Demand*, there was a significant effect of CONDITION (Figure 6.10c) and post hoc tests showed that NONE (MDN = 61, IQR = 47) and FEEDBACK (MDN = 70.5, IQR = 31) were associated with higher scores than PROMPT (MDN = 40.5, IQR = 52.25), BLANK (MDN = 20.5, IQR = 30.75), and SUMMARY (MDN = 11.5, IQR = 22.25).

There was also a significant effect of CONDITION on *Performance* (Figure 6.10d), with post hoc tests revealing that participants generally felt like they did better at the task with BLANK (MDN = 100, IQR = 8.5) than NONE (MDN = 90.5, IQR = 23.5), FEEDBACK (MDN = 95, IQR = 15.25), and SUMMARY (MDN = 100, IQR = 20.75).

For *Effort*, a significant effect of CONDITION (Figure 6.10e) and post hoc tests showed that NONE (MDN = 62.5, IQR = 42.75) and FEEDBACK (MDN = 68,

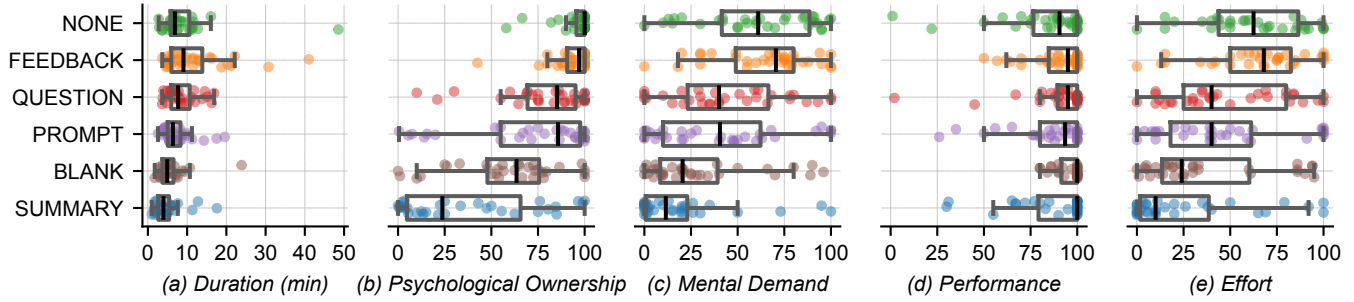


Figure 6.10: Experiment 3 results: (a) *Duration*, (b) *Psychological Ownership*, (c) *Mental Demand*, (d) *Performance*, and (e) *Effort*.

IQR = 32.5) had the highest scores, and SUMMARY (MDN = 10, IQR = 36.75) had the lowest. We did not observe significant differences between the different techniques for *Physical Demand*, *Temporal Demand*, and *Frustration*.

6.8.2.5 Preferences

There were no differences in *Frequency of Use*, with all medians being 50 or greater, suggesting that participants would like to use all techniques. For *Overall Ranking*, the Condorcet voting method revealed that overall, BLANK was ranked first, followed by QUESTION, SUMMARY, PROMPT, FEEDBACK, and NONE. This *Overall Ranking* suggests that participants generally preferred techniques with more AI involvement (BLANK, QUESTION, SUMMARY, and PROMPT) over those with more human involvement (FEEDBACK and NONE), for example: “the more involved AI was, the more I appreciated the help” (P15).

Considering the *Ranking* scores, 21 (66%) ranked BLANK within the top 3, and 20 (62%) placed QUESTION within the top 3. These two techniques were generally valued as they “had the greatest pedagogical heft” (P7) and “actively engage memory and reinforce understanding” (P4). Nineteen (59%) placed SUMMARY within the top 3, however, eight (25%) gave it a rank of 6, suggesting that opinions were more divided for SUMMARY. Those that ranked SUMMARY highly generally valued it for “[making] work so much easier” (P9), however, some noted how they “want to do things [and] feel accomplished” (P30), goals that were not as well-supported with SUMMARY. Fifteen (47%) placed PROMPT within the top 3 as it was “simple and easy” (P23) and “encouraged curiosity” (P4). In contrast, seventeen (53%) ranked FEEDBACK within the bottom 3, and 19 (59%) ranked NONE within the bottom 3. Both techniques required “too much effort” (P20), which could be “draining and mind-absorbing” (P25).

6.8.3 Summary

These results did not find differences in reading comprehension, even though many techniques were more mentally demanding and effortful. Participants generally preferred techniques with more AI involvement, even though these techniques were associated with less psychological ownership.

6.9 DISCUSSION

Overall, our results show that participants preferred AI margin notes over traditional, chat-based interactions. When AI margin notes were created through manual text selection, participants took longer to read the document and this required more effort, but they felt more psychological ownership and felt like they did a better job as a result. These feelings may have contributed to the strong preferences they had for manually-created AI margin notes. In short, our results suggest that AI margin notes were valued for their integration, and are perceived to be especially valuable with manual text selection.

Regarding human and AI involvement, our results suggest that different techniques have their own strengths and weaknesses. For example, techniques that required more human involvement, like receiving feedback about written text, required more mental demand and effort, but were associated with higher feelings of psychological ownership than those with low human involvement, like receiving summaries. However, participants preferred techniques with more AI involvement. One promising technique was receiving generated summaries with fill-in-the-blank exercises, the most preferred technique. It was perceived as not too mentally demanding or effortful, still associated with moderate feelings of psychological ownership, and participants felt like they performed well when creating them.

The implications for design are *to adopt the integrated AI margin note approach within document reader software, to support manual text selection, and to provide options of AI margin note technique that vary in human and AI involvement, depending on the user's preferences and goals.*

6.9.1 Reading Comprehension

Given literature suggests that increased cognitive engagement can increase learning, we were surprised to see that more cognitively demanding AI margin note techniques were not associated with higher reading comprehension scores. To confirm this is not due to one specific experimental protocol, we conducted other pilot experiments with different participants and key variations.

For example, we ran experiments with 24 hour gaps between the reading and testing stages to ensure that the documents were not simply remembered in short-term memory, but we did not observe any significant differences. Participants self-declared their knowledge of the topics discussed in each document, so we tried omitting participants with higher background knowledge, and found similar results. Another possibility is a ceiling effect because the range of possible scores is low (0-6). To rule this out, we conducted another pilot experiment using documents and questions from Guidroz et al. [69],³ enabling more granular (0-12) scores. These documents were also more difficult (university graduate level, rather than an 8th grade level). Even with this increased score granularity and more difficult reading task, we found similar results, with no significant differences between techniques.

³ We emailed the authors and received permission to reuse their materials.

Therefore, we believe that less cognitively engaging techniques with more AI involvement, like receiving generated summaries, may not be as detrimental to reading comprehension as one may assume.

6.9.2 Conflicting User Preferences Regarding AI Involvement

The concept of “manual” and “automatic” text selection explored in Experiment 2 can also be thought of as human and AI involvement, with manual text selection requiring more human involvement than automatic text selection. When reframed like this, we notice that the results from Experiments 2 and 3 share many similarities: both suggested that techniques with more human involvement are slower and more effortful, but associated with more psychological ownership. Yet the results revealed differing user preferences, with participants preferring less AI involvement when selecting text but more AI involvement in the actual comment.

These results may feel conflicting at first. But together, they provide additional insight into factors that are more important to readers when using LLMs to read and take notes. Based on prior work on text highlighting [192] and our findings in Chapter 4, we hypothesize that manual selection forces the reader to identify text they wish to learn more about, which requires them to recognize gaps in their own knowledge or understanding of the text. An LLM does not know what these gaps are, making it less capable of automatically positioning AI margin notes in ways that align with the reader’s needs. Identifying the right text to select is especially important, as the selected text also acts as contextual information for content that is produced and displayed within the comment. As such, selection automation may be a task that is less compatible with more AI involvement, resulting in a stronger preference for manual selection. However, once the AI margin note is correctly positioned, users may see more potential for AI (or perhaps see less issues associated with its use), and prefer techniques with more AI involvement as a result.

6.9.3 Other AI Margin Note Designs

We focused primarily on creating AI margin notes based on typed prompts and text responses. However, AI margin notes could also work well with other types of prompts and responses. For example, margin notes can also be hand-written using pen input [150]. Recent work has even shown that hand-written notes and sketches can be used to prompt LLMs [188], so AI margin notes could also respond to hand-written prompts. Prior work also suggests that interactive visual elements, like simulations [70] and charts [117] created for individual documents can further augment reading and learning. LLMs are capable of producing interactive visual elements like these, which could be integrated into AI margin notes. For example, if the reader selects text that describes different parts of flowering plants, the fill-in-the-blank technique could be augmented by showing a diagram and requiring readers to label it with the correct parts.

We focused on the specific context of creating AI margin notes as a study aid. However, people leave comments in document readers for many other reasons, for example, to leave editorial notes. Other AI margin note designs may be better suited to these goals. For example, after selecting text and writing a prompt like “improve the grammar of this sentence,” the AI margin note could show a modified version of the sentence underneath the prompt. To integrate these changes into the main text, the user must press a button to “accept” this change, much like the “Suggestion” feature of Google Docs.

6.9.4 *Limitations*

Our results, especially results related to reading comprehension, may not hold after extended long-term use. For example, LLMs may hallucinate and produce incorrect AI margin notes, and people may come to over-rely by studying incorrect notes [177]. Techniques with low human involvement may make students ‘lazy’ over time and impact their ability to take notes in situations where LLMs are unavailable.

Our results may not apply to different types of documents and longer documents. For example, as discussed in [Chapter 4](#), prior work on text highlighting suggests that readers should be able to distinguish between important and unimportant text [60], however, readers may struggle to identify the most important information within a long and complex document about an unfamiliar topic. In these cases, automatically selecting text within the document to create AI margin notes may be preferred.

Some aspects of our experimental design may be lacking in ecological validity. Notably, we chose to keep the number of AI margin notes and chat responses constant across techniques, and participants could only interact with one technique at a time. We made these decisions for increased control, however, such restrictions would not exist in a real system, where users would create as many AI margin notes and in whatever way they wish.

6.10 CONCLUSION AND FUTURE WORK

We propose and explore the design of “AI margin notes” that leverage the commenting feature of document reader software to provide LLM capabilities in a way that is more integrated into document text. Three experiments evaluated variations from different design parameters: integration, selection automation, and human and AI involvement, and overall, participants valued having integrated AI margin notes and creating them manually. AI margin note techniques that involved the human and AI to different degrees were valued for different reasons, suggesting that document reader software should provide multiple variations to support different user goals and preferences.

Future work should incorporate AI margin notes into existing software, explore other types of AI margin notes, and investigate the effects on factors like reading comprehension over the long term and with different types of documents. Our work adds more evidence that chat-based interfaces are not the only way of interacting with LLMs, and that increased integration with

document text is beneficial, especially when they are created manually and the trade-offs of human and AI involvement are considered.

CONCLUSION AND FUTURE DIRECTIONS

The goal of this dissertation was to:

Evaluate the effects of non-error-proofing user interface constraints while reading and writing, and design interaction techniques that use ‘soft’ constraints.

To achieve this goal, I proposed constraining *where* and *when* interaction could occur during different reading and writing activities. Here, I summarize the contributions of our work, implications for design, and opportunities for future work.

7.1 SUMMARY OF CONTRIBUTIONS

This dissertation primarily contributes knowledge that helps us understand the types of reading and writing activities that would benefit from constrained interactions, and the types of constraints that are effective in doing so. A secondary contribution is knowledge learned from designing novel interaction techniques that leverage constraints in ways that go beyond error-proofing. According to the taxonomy of HCI contributions proposed by Wobbrock and Kientz [182], these would correspond to empirical research contributions and artifact contributions.

7.1.1 Empirical Research Contributions

Each project consisted of one or more controlled experiment that aimed to better understand whether interactions with ‘hard’ constraints would improve user outcomes. It is only through these experiments we were able to learn that scrolling may not hinder reading comprehension, as was previously believed; word limits on highlighting can improve reading comprehension; psychological ownership increases with longer prompts; and positioning prompts beside document text is desirable. These findings expand our knowledge of the use cases for, and the benefits that arise from constraining user interactions. Notably, each experiment had high internal validity as every experimental design controlled for specific factors we wished to test, yet also had high external validity as our experimental interfaces reproduced real and commonly-used interfaces and our sample sizes were large and diverse, which is surprisingly rare for HCI research [31]. Of course, our studies may be lacking in ecological validity as we evaluated reading and writing specific types of documents using participants in North America, but given the context we studied, these experiments were conducted with high rigour. All of our experimental data is (or will soon be) publicly-available on the ACM Digital Library.

		Constraint Type				Constraint Goal
		Where	When	Who	What	Why
User Activity	Reading	CHAPTER 3	CHAPTER 4	Future Work		
	Writing	Future Work	CHAPTER 5			
	Reading + Writing	CHAPTER 6				

Table 7.1: The design space explored in this dissertation could be further expanded. Grey, italicized text indicates new additions, and blue indicates potential avenues for future work.

7.1.2 Artifact Contributions

When constrained interactions showed promise, we created interaction techniques using ‘soft’ constraints to introduce friction. We proposed making highlighting more temporally, physically, and perceptually more difficult to discourage excessive highlighting for improved reading comprehension; we proposed making it more temporally and physically demanding to submit short prompts for increased psychological ownership; and we created new ways for readers to create margin notes with AI, many of which had requirements that had to be satisfied to continue. Creating these interaction techniques enabled more empirical research contributions. Through additional pilots and user studies, we discovered that ‘soft’ constraints may not work as well as ‘hard’ constraints when highlighting, but work especially well to increase psychological ownership when prompting with traditional chat-based interfaces and when writing AI margin notes. These interaction techniques can be easily integrated into new and existing reading and writing systems, making them valuable for others to build upon.

7.2 EFFECTIVENESS OF CONSTRAINTS FOR READING AND WRITING

Returning to the design space first presented in [Chapter 1](#), this dissertation explored how *where* and *when* interaction constraints affect different *reading* and *writing* activities ([Table 7.1](#)). It builds on prior work in two primary ways. First, it considers a more diverse and representative set of activities that have not been explored before, for example, [active reading](#) by highlighting and taking notes, and writing with emerging technology like LLMs. Second, it considers a wider range of constraints, such as word limits to restrict *when* interaction can occur and anchoring interface elements to restrict *where* interaction can occur. Overall, our results suggest that *where* constraints can benefit prompting, but may not benefit [passive reading](#) on a phone. However, all ‘hard’ *when* constraints seemed to benefit all *reading* and *writing* activities. This suggests that ‘hard’ *when* constraints may be especially effective for HCI researchers and practitioners to consider when designing new constrained reading and writing interfaces.

7.3 DIFFICULTIES WITH DESIGNING NEW INTERACTION TECHNIQUES

Although constraining *when* interaction could occur seemed to benefit *reading* and *writing*, we did not have similar success when trying to integrate ‘soft’ constraints into interaction techniques. Specifically, when we tried to make highlighting more difficult, temporal and perceptual costs did not seem to change how many words participants highlighted. Even when motor costs did, the reductions in word count were not associated with improved reading comprehension. This is surprising, as similar types of mechanics in the form of temporal and motor costs were introduced in [Chapter 5](#) yet they were both effective at nudging users.

As discussed in [Chapter 4](#), it is possible that the constrained highlighting interaction techniques need to be further tuned to be effective. However, another possibility is that the successful techniques proposed to increase prompt length made the concept of *choice* more explicit to users. Specifically, the techniques required users to fill a ‘gauge’ in one of two ways: writing more words or spending more time or exerting more physical effort to submit shorter prompts. In contrast, the highlighting techniques did not provide multiple ways of highlighting: users had to highlight text using the same technique throughout the entire document. Although we anticipated that experiencing more extreme time delays, circle diameters, and obscure percents would cause users to shift from System 1 to System 2 thinking and change their highlighting behaviour accordingly ([dual process theory](#)), our participants may not have realized that this was a possibility due to the lack of alternatives to choose from. As a result, they may have continued to persevere through this difficult highlighting task to do it ‘well’ due to high conscientiousness [52] and perhaps pressure from the crowdsourcing experiment service.

Furthermore, these constrained highlighting techniques were surfaced *after* the user decided that they would like to highlight specific text, and encouraging users to deviate from this goal after it has been set may be more difficult. But having a ‘hard’ constraint on highlighting and even visualizing a progress bar makes the idea of choice and goal-setting more explicit *before* they begin using the system, so they may be more likely to adjust their behaviours. Therefore, it may be more effective to surface such *when* constraints earlier, and potentially use feedforward to make the concept of choice more obvious within the interface.

To better understand factors that contributed to the success of constrained interaction techniques, I plan on conducting more qualitative and process-oriented research in the future. For example, through in-person observational studies with think-aloud protocols and semi-structured interviews.

7.3.1 *Need for Design Guidelines*

Moving beyond these two projects, the challenges we experienced and the several design decisions we had to make demonstrate the need for *design guidelines that the HCI community can leverage for constrained interaction techniques*. For example, there are several choices of limits, ways of surfacing and mapping such limits to changes in the interface, and adapting them to

individual users and different contexts of use, like as user goals change, as expertise improves, non-experimental contexts, and short- and long-term use.

An ongoing goal of mine is to formalize the design of constrained interaction techniques and interfaces. More specifically, I aim to conduct several smaller, short-term experiments using prototype systems that implement variations of constrained interactions and develop systems that use constrained interactions to facilitate long-term evaluations. Most novel interaction techniques in HCI are not adopted broadly or integrated into real-world systems, perhaps because of a lack of guiding principles and rules that help designers integrate them into systems [7, 8], making this an important next step.

As part of these design guidelines, it may even be necessary to rethink the constraint types originally proposed by Norman [128, p. 125-132]. For example, the idea of a cursor being physically constrained to a screen [129] does not apply to new augmented reality input techniques that move the cursor into the real world [92]. It may also be necessary to formalize other constraint types that better represent the specific ways in which constraints were used for non-error-proofing purposes throughout this dissertation, and to create a unique name for such “constrained interaction techniques.”

I really like “bounded interactions” as a name for such “constrained interaction techniques.” Thanks, Takeo Igarashi, for suggesting this name during the UIST 2023 Doctoral Symposium!

7.4 OTHER POSSIBILITIES FOR CONSTRAINED READING AND WRITING

Our design space focused on a subset of constraints that restrict *when* and *where* interaction can occur when reading and writing, but there are several other possibilities. For example, one possibility is to build on previous work about time-bounded reading [38, 174]. Systems could be built to adaptively estimate reading duration for any given document and use this to restrict *when* reading can occur for more focus and comprehension. Another possibility is to restrict *where* highlighting can occur in a document by only allowing key words and phrases to be highlighted (i.e., non-relevant text cannot be selected).

Notably, this dissertation did not explore restricting *where* writing can take place, but future work could explore this idea. For example, free-form editorial notes (either written using the comment feature of word processors or with a digital pen) could be processed to create action items tied to specific parts of the text. In a word processor, the writer could enter an “editorial mode” where they are presented with a single action item and are only permitted to edit the corresponding text. Constraining where writing can take place in this way may increase focus and editing speed.

Our work also focused on reading and writing activities for a single user, but constraints may also be beneficial in collaborative, multi-user environments, especially for writing. For example, in a related project not appearing in this dissertation, we proposed delaying *when* edits are shared with others in synchronous collaborative writing tools like Google Docs to help writers feel more comfortable [187]. Many non-real-time techniques, like delaying updates to when entire sentences were typed, were perceived to be desirable and beneficial for increasing comfort while writing. However, building constrained interaction techniques for such collaborative environments introduces several technical and design challenges. For example, it is unclear if a system should impose different types or levels of constraints for each user, or

if these should be applied consistently or shared across all users. Depending on the design, managing and synchronizing constraint states across users could be challenging, especially if different users have differing constraint states that conflict with each other.

Our design space could also be expanded to make the constraint goal more explicit (i.e., *why*; Table 7.1). For example, we proposed increasing prompt length to improve psychological ownership, but the experimental protocol could have been adapted to improve learning outcomes by increasing cognitive effort [20–22] and reducing over-reliance [30]. Additional *Constraint Types* could be explored too. Prior work suggests that restricting *who* can access documents may help writers feel more comfortable with sharing work-in-progress earlier [16, 100]. Thinking about *what* interactions are permitted to achieve a goal provides another way to think about AI margin note techniques. Some only allowed certain types of input, such as typing text or selecting items from a drop down. When reading, one new idea is to only create highlights if the selected text only consists of key words and phrases, constraining *what* highlights can consist of, much like how some text fields for phone numbers prevent the user from typing non-numerical characters.

7.5 BEYOND READING AND WRITING

Future work could also explore other applications of constraints that go beyond reading and writing, building on prior work in HCI that has used non-error-proofing constraints (Section 2.2). For example, as previously mentioned, I created MicroMentor, a help system constrained the duration of help sessions between novice and expert software users to three minutes. MicroMentor was valued for reducing the perceived transaction cost of asking for and providing help.

Recently, we explored making the act of taking photos more difficult by zooming in the viewfinder or requiring the user to describe objects being photographed to improve recall. Constraints could also be integrated into camera applications to limit how many photos and the types of photos people can take to improve attachment, or to create counterfunctional camera designs [137] that provoke thought or reflection. For example, preventing an item from being photographed if it is a certain colour or if it has already been photographed before.

Another idea is purposely limiting which commands are available in a tutorial to help users discover different features and new ways of using existing features. For example, Renom et al. [147] showed that in a text-editing context, disabling specific commands encourages users to re-purpose other commands, like using tabs and spaces to align text when ruler features are unavailable. Applying this concept to tutorial systems more broadly could help users learn more commands and potentially encourage them to overcome the “paradox of the active user,” the tendency to use known yet inefficient methods to accomplish a task over learning and trying new and more efficient methods [33].

7.6 WHO SHOULD IMPOSE CONSTRAINTS?

To explore the effectiveness of constraints when reading and writing in a more controlled way, we imposed mandatory constraints on users. My intention is to use constraints in ways that positively benefit users, however, there is the risk that constraints could be used by designers to guide users towards negative or self-serving outcomes (“dark patterns”). For example, requiring a user to press and hold a button or to ‘pump’ a slider to cancel an online subscription is problematic and unethical, however, these techniques may help users make fewer impulsive purchases by giving them more time to reflect. This further necessitates design guidelines that define when constraints are permitted to be used in non-error-proofing ways.

To mitigate these issues, it may be more appropriate for these constraints to be *self-imposed*, by allowing users to opt in to, and customize them as needed. Prior work suggests that many users are interested in self-imposed limits (e.g., [43, 93, 94]), however, it is also well-known that people do not deviate from default settings [82]. Perhaps a more realistic balance is to *reify* [9] user interface constraints by presenting them as tools that can be manipulated (i.e., instrumental interaction [7]), rather than as settings hidden away in menus.

7.7 FINAL WORD

Constraints are fundamental to human life and have been an important part of human-centred design for centuries. But perhaps ironically, we have been too limited by *how* we have applied constraints to user interface design, which has mostly focused on error-proofing. I believe that we can do more with constraints by putting them at the forefront of user interactions, especially for fundamental activities like reading and writing.

In this dissertation, we explored constraining *when* and *where* interaction can take place while reading and writing. We proposed constraining *where* the viewport can be placed when reading documents; *when* highlighting can take place; *when* prompts can be submitted to a generative AI assistant; *where* prompting can take place to read and write notes for a document; and *when* written AI margin notes could be finalized. We also designed and evaluated interaction techniques that use ‘soft’ constraints to give users more flexibility. Although such restrictions on *when* and *where* interaction could happen were not always successful, *they often can be*: they can increase reading comprehension when highlighting text and increase psychological ownership when writing short stories and when writing notes using LLMs. I believe that this application of user interface constraints has the potential to transform reading, writing, and more; and prioritizing constraints in design can significantly improve how we interact with user interfaces.

BIBLIOGRAPHY

- [1] Oguz A. Acar, Murat Tarakci, and Daan van Knippenberg. "Creativity and Innovation Under Constraints: A Cross-Disciplinary Integrative Review." In: *Journal of Management* 45.1 (2019), pp. 96–121. DOI: [10.1177/0149206318805832](https://doi.org/10.1177/0149206318805832).
- [2] Annette Adler, Anuj Gujar, Beverly L. Harrison, Kenton O'Hara, and Abigail Sellen. "A Diary Study of Work-Related Reading: Design Implications for Digital Reading Devices." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '98. Los Angeles, California, USA: ACM Press/Addison-Wesley Publishing Co., 1998, 241–248. ISBN: 0201309874. DOI: [10.1145/274644.274679](https://doi.org/10.1145/274644.274679).
- [3] Adobe. *Generative AI in Adobe Document Cloud Applications*. <https://helpx.adobe.com/ca/acrobat/using/generative-ai.html>. [Accessed 20-08-2025]. 2025.
- [4] Julie Alonzo, Gerald Tindal, Kirt Ulmer, and Aaron Glasgow. *easyCBM® online progress monitoring assessment system*. Eugene, OR: University of Oregon, Behavioral Research and Teaching. 2006. URL: <http://easycbm.com>.
- [5] Amid Ayobi, Tobias Sonne, Paul Marshall, and Anna L. Cox. "Flexible and Mindful Self-Tracking: Design Implications from Paper Bullet Journals." In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal, QC, Canada: Association for Computing Machinery, 2018, 1–14. ISBN: 9781450356206. DOI: [10.1145/3173574.3173602](https://doi.org/10.1145/3173574.3173602).
- [6] Markus Baer and Greg R. Oldham. "The Curvilinear Relation Between Experienced Creative Time Pressure and Creativity: Moderating Effects of Openness to Experience and Support for Creativity." In: *Journal of Applied Psychology* 91.4 (2006), p. 963. DOI: <https://doi.org/10.1037/0021-9010.91.4.963>.
- [7] Michel Beaudouin-Lafon. "Instrumental Interaction: An Interaction Model for Designing Post-WIMP User Interfaces." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '00. The Hague, The Netherlands: Association for Computing Machinery, 2000, 446–453. ISBN: 1581132166. DOI: [10.1145/332040.332473](https://doi.org/10.1145/332040.332473).
- [8] Michel Beaudouin-Lafon, Susanne Bødker, and Wendy E. Mackay. "Generative Theories of Interaction." In: *ACM Trans. Comput.-Hum. Interact.* 28.6 (Nov. 2021). ISSN: 1073-0516. DOI: [10.1145/3468505](https://doi.org/10.1145/3468505).
- [9] Michel Beaudouin-Lafon and Wendy E. Mackay. "Reification, Polymorphism and Reuse: Three Principles for Designing Visual Interfaces." In: *Proceedings of the Working Conference on Advanced Visual Interfaces*. AVI '00. Palermo, Italy: Association for Computing Machinery, 2000, 102–109. ISBN: 1581132522. DOI: [10.1145/345513.345267](https://doi.org/10.1145/345513.345267).

- [10] Kenneth E. Bell and John E. Limber. "Reading Skill, Textbook Marking, and Course Performance." In: *Literacy Research and Instruction* 49.1 (2009), pp. 56–67. DOI: [10.1080/19388070802695879](https://doi.org/10.1080/19388070802695879).
- [11] Luca Benedetti, Holger Winnemöller, Massimiliano Corsini, and Roberto Scopigno. "Painting with Bob: Assisted Creativity for Novices." In: *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. UIST '14. Honolulu, Hawaii, USA: Association for Computing Machinery, 2014, 419–428. ISBN: 9781450330695. DOI: [10.1145/2642918.2647415](https://doi.org/10.1145/2642918.2647415).
- [12] Simone Benedetto, Andrea Carbone, Marco Pedrotti, Kevin Le Fevre, Linda Amel Yahia Bey, and Thierry Baccino. "Rapid serial visual presentation in reading: The case of Spritz." In: *Computers in Human Behavior* 45 (2015), pp. 352–358. ISSN: 0747-5632. DOI: <https://doi.org/10.1016/j.chb.2014.12.043>.
- [13] Ben Bengler and Nick Bryan-Kinns. "Designing collaborative musical experiences for broad audiences." In: *Proceedings of the 9th ACM Conference on Creativity & Cognition*. C&C '13. Sydney, Australia: Association for Computing Machinery, 2013, 234–242. ISBN: 9781450321501. DOI: [10.1145/2466627.2466633](https://doi.org/10.1145/2466627.2466633).
- [14] Eric A. Bier and Maureen C. Stone. "Snap-dragging." In: *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '86. New York, NY, USA: Association for Computing Machinery, 1986, 233–240. ISBN: 0897911962. DOI: [10.1145/15922.15912](https://doi.org/10.1145/15922.15912).
- [15] Oloff C. Biermann, Ning F. Ma, and Dongwook Yoon. "From Tool to Companion: Storywriters Want AI Writers to Respect Their Personal Values and Writing Strategies." In: *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. DIS '22. Virtual Event, Australia: Association for Computing Machinery, 2022, 1209–1227. ISBN: 9781450393584. DOI: [10.1145/3532106.3533506](https://doi.org/10.1145/3532106.3533506).
- [16] Katherine Bindley. "The Five Most Dreaded Words at the Office: 'Let's Start a Google Doc'." Shared documents are a workplace minefield of performance anxiety and awkward interactions." In: *The Wall Street Journal* (2022). URL: <https://www.wsj.com/articles/google-doc-sharing-work-collaboration-11654612774>.
- [17] Jeremy Birnholtz, Stephanie Steinhardt, and Antonella Pavese. "Write Here, Write Now!: An Experimental Study of Group Maintenance in Collaborative Writing." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. Paris, France: Association for Computing Machinery, 2013, 961–970. ISBN: 9781450318990. DOI: [10.1145/2470654.2466123](https://doi.org/10.1145/2470654.2466123).
- [18] Michael Mose Biskjaer, Bo T. Christensen, Morten Friis-Olivarius, Sille J.J. Abildgaard, Caroline Lundqvist, and Kim Halskov. "How task constraints affect inspiration search strategies." In: *International Journal of Technology and Design Education* 30.1 (2020), pp. 101–125. DOI: <https://doi.org/10.1007/s10798-019-09496-7>.

- [19] Michael Mose Biskjaer, Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, and Peter Dalsgaard. "How Time Constraints in a Creativity Support Tool Affect the Creative Writing Experience." In: *Proceedings of the 31st European Conference on Cognitive Ergonomics*. ECCE '19. Belfast, United Kingdom: Association for Computing Machinery, 2019, 100–107. ISBN: 9781450371667. DOI: [10.1145/3335082.3335084](https://doi.org/10.1145/3335082.3335084).
- [20] Elizabeth L. Bjork and Robert A. Bjork. "Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning." In: *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* 2.59-68 (2011), pp. 56–64.
- [21] Robert A. Bjork. "Institutional Impediments to Effective Training." In: *Learning, Remembering, Believing: Enhancing Human Performance* (1994), pp. 295–306.
- [22] Robert A. Bjork. "Memory and Metamemory Considerations in the Training of Human Beings." In: *Metacognition: Knowing about Knowing* 185.7.2 (1994), pp. 185–205.
- [23] Robert A. Bjork. "Assessing Our Own Competence: Heuristics and Illusions." In: *Attention and Performance: Cognitive Regulation of Performance: Interaction of Theory and Application* 17 (1999), pp. 435–459.
- [24] Ina Blau and Avner Caspi. "Sharing and Collaborating with Google Docs: The Influence of Psychological Ownership, Responsibility, and Student's Attitudes on Outcome Quality." In: *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*. Association for the Advancement of Computing in Education (AACE). 2009, pp. 3329–3335.
- [25] Ina Blau and Avner Caspi. "What Type of Collaboration Helps? Psychological Ownership, Perceived Learning and Outcome Quality of Collaboration using Google Docs." In: *Proceedings of the Chais Conference on Instructional Technologies Research*. Vol. 12. 1. 2009, pp. 48–55.
- [26] Margaret A. Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2004.
- [27] Burke H. Bretzing and Raymond W. Kulhavy. "Notetaking and Depth of Processing." In: *Contemporary Educational Psychology* 4.2 (1979), pp. 145–153. ISSN: 0361-476X. DOI: [10.1016/0361-476X\(79\)90069-9](https://doi.org/10.1016/0361-476X(79)90069-9).
- [28] Burke H. Bretzing and Raymond W. Kulhavy. "Note-Taking and Passage Style." In: *Journal of Educational Psychology* 73.2 (1981), p. 242. DOI: [10.1037/0022-0663.73.2.242](https://doi.org/10.1037/0022-0663.73.2.242).
- [29] Zvia Breznitz and David L. Share. "Effects of Accelerated Reading Rate on Memory for Text." In: *Journal of Educational Psychology* 84.2 (1992), p. 193. DOI: [10.1037/0022-0663.84.2.193](https://doi.org/10.1037/0022-0663.84.2.193).
- [30] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making." In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: [10.1145/3449287](https://doi.org/10.1145/3449287).

- [31] Kelly Caine. "Local Standards for Sample Size at CHI." In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. San Jose, California, USA: Association for Computing Machinery, 2016, 981–992. ISBN: 9781450333627. DOI: [10.1145/2858036.2858498](https://doi.org/10.1145/2858036.2858498).
- [32] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. "23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–15. ISBN: 9781450359702. DOI: [10.1145/3290605.3300733](https://doi.org/10.1145/3290605.3300733).
- [33] John M. Carroll and Mary Beth Rosson. "Paradox of the active user." In: *Interfacing Thought: Cognitive Aspects of Human-Computer Interaction*. Cambridge, MA, USA: MIT Press, 1987, 80–111. ISBN: 0262031256.
- [34] Ryder Carroll. *The Bullet Journal Method: Track the Past, Order the Present, Design the Future*. Penguin, 2018.
- [35] Géry Casiez, Nicolas Roussel, and Daniel Vogel. "1€ Filter: A Simple Speed-based Low-pass Filter for Noisy Input in Interactive Systems." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA: Association for Computing Machinery, 2012, 2527–2530. ISBN: 9781450310154. DOI: [10.1145/2207676.2208639](https://doi.org/10.1145/2207676.2208639).
- [36] Avner Caspi and Ina Blau. "Collaboration and Psychological Ownership: How does the Tension between the Two Influence Perceived Learning?" In: *Social Psychology of Education* 14 (2011), pp. 283–298. DOI: [10.1007/s11218-010-9141-z](https://doi.org/10.1007/s11218-010-9141-z).
- [37] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. "Universal Sentence Encoder for English." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 169–174. DOI: [10.18653/v1/D18-2029](https://doi.org/10.18653/v1/D18-2029).
- [38] Anna Chang. "The Effect of a Timed Reading Activity on EFL Learners: Speed, Comprehension, and Perceptions." In: *Reading in a Foreign Language* 22.2 (2010), pp. 284–303.
- [39] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. "Supporting Mobile Sensemaking Through Intentionally Uncertain Highlighting." In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. UIST '16. Tokyo, Japan: Association for Computing Machinery, 2016, 61–68. ISBN: 9781450341899. DOI: [10.1145/2984511.2984538](https://doi.org/10.1145/2984511.2984538).
- [40] University of North Carolina at Chapel Hill. *Highlighting – Learning Center*. <https://learningcenter.unc.edu/tips-and-tools/using-highlighters/>. [Accessed 30-08-2023]. 2016.
- [41] ChatPDF. *ChatPDF AI | Chat with any PDF | Free*. <https://www.chatpdf.com/>. [Accessed 29-08-2025]. 2023.

- [42] Christy Choi and Francesca Annio. *The Winner of a Prestigious Japanese Literary Award has Confirmed AI Helped Write her Book*. 2024. URL: <https://www.ctvnews.ca/sci-tech/the-winner-of-a-prestigious-japanese-literary-award-has-confirmed-ai-helped-write-her-book-1.6734205>.
- [43] Francesco Cirillo. "The Pomodoro Technique (The Pomodoro)." In: *Agile Processes in Software Engineering and* 54.2 (2006), p. 35.
- [44] Andy Cockburn, Per Ola Kristensson, Jason Alexander, and Shumin Zhai. "Hard Lessons: Effort-Inducing Interfaces Benefit Spatial Learning." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. San Jose, California, USA: Association for Computing Machinery, 2007, 1571–1580. ISBN: 9781595935939. DOI: [10.1145/1240624.1240863](https://doi.org/10.1145/1240624.1240863).
- [45] Tim Coughlan and Peter Johnson. "Constrain Yourself: Exploring End User Development in Support for Musical Creativity." In: *Proceedings of the 6th ACM SIGCHI Conference on Creativity & Cognition*. C&C '07. Washington, DC, USA: Association for Computing Machinery, 2007, 247–248. ISBN: 9781595937124. DOI: [10.1145/1254960.1254996](https://doi.org/10.1145/1254960.1254996).
- [46] Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. "Design Frictions for Mindful Interactions: The Case for Microboundaries." In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA '16. San Jose, California, USA: Association for Computing Machinery, 2016, 1389–1397. ISBN: 9781450340823. DOI: [10.1145/2851581.2892410](https://doi.org/10.1145/2851581.2892410).
- [47] Fergus I.M. Craik and Robert S. Lockhart. "Levels of Processing: A Framework for Memory Research." In: *Journal of Verbal Learning and Verbal Behavior* 11.6 (1972), pp. 671–684. ISSN: 0022-5371. DOI: [10.1016/S0022-5371\(72\)80001-X](https://doi.org/10.1016/S0022-5371(72)80001-X).
- [48] Darren W. Dahl and C. Page Moreau. "Thinking Inside the Box: Why Consumers Enjoy Constrained Creative Experiences." In: *Journal of Marketing Research* 44.3 (2007), pp. 357–369. DOI: [10.1509/jmkr.44.3.357](https://doi.org/10.1509/jmkr.44.3.357).
- [49] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. "Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: [10.1145/3544548.3580672](https://doi.org/10.1145/3544548.3580672).
- [50] Ruiqi Deng, Maoli Jiang, Xinlu Yu, Yuyan Lu, and Shasha Liu. "Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies." In: *Computers & Education* 227 (2025), p. 105224. ISSN: 0360-1315. DOI: [10.1016/j.compedu.2024.105224](https://doi.org/10.1016/j.compedu.2024.105224).

- [51] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. "Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Models." In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: [10.1145/3613904.3642134](https://doi.org/10.1145/3613904.3642134).
- [52] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. "Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA." In: *PLOS ONE* 18.3 (Mar. 2023), pp. 1–17. DOI: [10.1371/journal.pone.0279720](https://doi.org/10.1371/journal.pone.0279720).
- [53] Steven P. Dow, Kate Heddlestone, and Scott R. Klemmer. "The Efficacy of Prototyping Under Time Constraints." In: *Proceedings of the Seventh ACM Conference on Creativity and Cognition*. C&C '09. Berkeley, California, USA: Association for Computing Machinery, 2009, 165–174. ISBN: 9781605588650. DOI: [10.1145/1640233.1640260](https://doi.org/10.1145/1640233.1640260).
- [54] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. "The AI Ghostwriter Effect: When Users do not Perceive Ownership of AI-Generated Text but Self-Declare as Authors." In: *ACM Trans. Comput.-Hum. Interact.* 31.2 (2024). ISSN: 1073-0516. DOI: [10.1145/3637875](https://doi.org/10.1145/3637875).
- [55] Mark D. Dunlop, Emma Nicol, Andreas Komninos, Prima Dona, and Naveen Durga. "Measuring Inviscid Text Entry using Image Description Tasks." In: *Proceedings of the 2016 CHI Workshop on Inviscid Text Entry and Beyond*. 2016.
- [56] Serge Egelman, David Molnar, Nicolas Christin, Alessandro Acquisti, Cormac Herley, and Shriram Krishnamurthi. "Please Continue to Hold: An empirical study on user tolerance of security delays." In: *Ninth Workshop on the Economics of Information Security*. 2010.
- [57] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. "An Aligned Rank Transform Procedure for Multifactor Contrast Tests." In: *The 34th Annual ACM Symposium on User Interface Software and Technology*. UIST '21. Virtual Event, USA: Association for Computing Machinery, 2021, 754–768. ISBN: 9781450386357. DOI: [10.1145/3472749.3474784](https://doi.org/10.1145/3472749.3474784).
- [58] Jon Elster. *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints*. New York, NY: Cambridge University Press, 2000.
- [59] Margaret Foley, Géry Casiez, and Daniel Vogel. "Comparing Smartphone Speech Recognition and Touchscreen Typing for Composition and Transcription." In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–11. ISBN: 9781450367080. DOI: [10.1145/3313831.3376861](https://doi.org/10.1145/3313831.3376861).
- [60] Robert L. Fowler and Anne S. Barker. "Effectiveness of Highlighting for Retention of Text Material." In: *Journal of Applied Psychology* 59.3 (1974), pp. 358–364. DOI: [10.1037/h0036750](https://doi.org/10.1037/h0036750).

- [61] Takugo Y. Fukaya, Susumu Ono, Minoru Minakuchi, Seiya Nakashima, Masako Hayashi, and Hiroshi Ando. "Reading Text on a Smart Phone: Scrolling vs. Paging: Toward Designing Effective Electronic Manuals." In: *2011 International Conference on User Science and Engineering (i-USER)*. 2011, pp. 59–63. DOI: [10.1109/iUSER.2011.6150537](https://doi.org/10.1109/iUSER.2011.6150537).
- [62] Katy Ilonka Gero and Lydia B. Chilton. "Metaphoria: An Algorithmic Companion for Metaphor Creation." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–12. ISBN: 9781450359702. DOI: [10.1145/3290605.3300526](https://doi.org/10.1145/3290605.3300526).
- [63] James Jerome Gibson. *The Ecological Approach to Visual Perception*. Boston: Houghton, Mifflin and Company, 1979.
- [64] Google. *Google NotebookLM | AI Research Tool & Thinking Partner*. <https://notebooklm.google>. [Accessed 20-08-2025]. 2025.
- [65] Scott C. Griffith. "Cooperative Learning Techniques in the Classroom." In: *Journal of Experiential Education* 13.2 (1990), pp. 41–44. DOI: [10.1177/105382599001300208](https://doi.org/10.1177/105382599001300208).
- [66] Tim Groot Kormelink and Irene Costera Meijer. "Material and sensory dimensions of everyday news use." In: *Media, Culture & Society* 41.5 (2019), pp. 637–653. DOI: [10.1177/0163443718810910](https://doi.org/10.1177/0163443718810910).
- [67] Axel Grund, Stefan Fries, Matthias Nückles, Alexander Renkl, and Julian Roelle. "When is Learning "Effortful"? Scrutinizing the Concept of Mental Effort in Cognitively Oriented Research from a Motivational Perspective." In: *Educational Psychology Review* 36.1 (2024), p. 11. DOI: [10.1007/s10648-024-09852-7](https://doi.org/10.1007/s10648-024-09852-7).
- [68] Ziwei Gu, Ian Arawjo, Kenneth Li, Jonathan K. Kummerfeld, and Elena L. Glassman. "An AI-Resilient Text Rendering Technique for Reading and Skimming Documents." In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: [10.1145/3613904.3642699](https://doi.org/10.1145/3613904.3642699).
- [69] Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias MJ Bellaiche, Miguel Ángel Garrido, Faruk Ahmed, Divyansh Choudhary, Jay Hartford, Chenwei Xu, Henry Javier Serrano Echeverria, Yifan Wang, Jeff Shaffer, Eric, Cao, Yossi Matias, Avinatan Hassidim, Dale R Webster, Yun Liu, Sho Fujiwara, Peggy Bui, and Quang Duong. *LLM-based Text Simplification and its Effect on User Comprehension and Cognitive Load*. 2025. arXiv: [2505.01980](https://arxiv.org/abs/2505.01980) [cs.CL].
- [70] Aditya Gunturu, Yi Wen, Nandi Zhang, Jarin Thundathil, Rubaiat Habib Kazi, and Ryo Suzuki. "Augmented Physics: Creating Interactive and Embedded Physics Simulations from Static Textbook Diagrams." In: *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. UIST '24. Pittsburgh, PA, USA: Association for Computing Machinery, 2024. ISBN: 9798400706288. DOI: [10.1145/3654777.3676392](https://doi.org/10.1145/3654777.3676392).

- [71] Aaron Halfaker, Aniket Kittur, Robert Kraut, and John Riedl. "A Jury of your Peers: Quality, Experience and Ownership in Wikipedia." In: *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. WikiSym '09. Orlando, Florida: Association for Computing Machinery, 2009. ISBN: 9781605587301. DOI: [10.1145/1641309.1641332](https://doi.org/10.1145/1641309.1641332).
- [72] Han L. Han, Miguel A. Renom, Wendy E. Mackay, and Michel Beaudouin-Lafon. "Textlets: Supporting Constraints and Consistency in Text Documents." In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–13. ISBN: 9781450367080. DOI: [10.1145/3313831.3376804](https://doi.org/10.1145/3313831.3376804).
- [73] Beverly L. Harrison, Hiroshi Ishii, Kim J. Vicente, and William A. S. Buxton. "Transparent Layered User Interfaces: An Evaluation of a Display Design to Enhance Focused and Divided Attention." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '95. Denver, Colorado, USA: ACM Press/Addison-Wesley Publishing Co., 1995, 317–324. ISBN: 0201847051. DOI: [10.1145/223904.223945](https://doi.org/10.1145/223904.223945).
- [74] Catrinel Haught and Philip N. Johnson-Laird. "Creativity and Constraints: The Production of Novel Sentences." In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 25. 25. 2003.
- [75] Ymkje E. Haverkamp, Ivar Bråten, Natalia Latini, and Ladislao Salmerón. "Is it the size, the movement, or both? Investigating effects of screen size and text movement on processing, understanding, and motivation when students read informational text." In: *Reading and Writing* 36.7 (2022), pp. 1589–1608. DOI: [10.1007/s11145-022-10328-9](https://doi.org/10.1007/s11145-022-10328-9).
- [76] Shih-Wen Hsiao, Ming-Feng Wang, and Chien-Wie Chen. "Time pressure and creativity in industrial design." In: *International Journal of Technology and Design Education* 27.2 (2017), pp. 271–289. DOI: [10.1007/s10798-015-9343-y](https://doi.org/10.1007/s10798-015-9343-y).
- [77] Yao-Hung Hsieh, Chiuhsiang Joe Lin, Hsiao-Ching Chen, Ting-Ting Huang, and James C. Chen. "Reading Performance of Chinese Text with Automatic Scrolling." In: *Human-Computer Interaction. Interaction Platforms and Techniques*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 311–319. ISBN: 978-3-540-73107-8.
- [78] Chieh-Yang Huang, Sanjana Gautam, Shannon McClellan Brooks, Ya-Fang Lin, Tiffany Kneareem, and Ting-Hao 'Kenneth' Huang. *Inspo: Writing with Crowds Alongside AI*. 2024. arXiv: [2311.16521](https://arxiv.org/abs/2311.16521) [cs.HC].
- [79] Nanna Inie, Jonas Frich, and Peter Dalsgaard. "How Researchers Manage Ideas." In: *Proceedings of the 14th Conference on Creativity and Cognition*. C&C '22. Venice, Italy: Association for Computing Machinery, 2022, 83–96. ISBN: 9781450393270. DOI: [10.1145/3527927.3532813](https://doi.org/10.1145/3527927.3532813).
- [80] Heather Joanna Jackson. *Marginalia: Readers Writing in Books*. Yale University Press, 2001.

- [81] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. “Why we Twitter: Understanding Microblogging Usage and Communities.” In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*. WebKDD/SNA-KDD ’07. San Jose, California: Association for Computing Machinery, 2007, 56–65. ISBN: 9781595938480. DOI: [10.1145/1348549.1348556](https://doi.org/10.1145/1348549.1348556).
- [82] Eric J. Johnson and Daniel Goldstein. “Do Defaults Save Lives?” In: *Science* 302.5649 (2003), pp. 1338–1339. DOI: [10.1126/science.1091721](https://doi.org/10.1126/science.1091721).
- [83] Nikhita Joshi, Richard Li, Jiannan Li, Leonardo Pavanatto, Michel Pahud, Jatin Sharma, Bongshin Lee, Hugo Romat, William Buxton, Nicolai Marquardt, Ken Hinckley, and Nathalie Henry Riche. “Opportunistic Nudges for Task Migration Between Personal Devices.” In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. CHI EA ’24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703317. DOI: [10.1145/3613905.3651037](https://doi.org/10.1145/3613905.3651037).
- [84] Nikhita Joshi, Justin Matejka, Fraser Anderson, Tovi Grossman, and George Fitzmaurice. “MicroMentor: Peer-to-Peer Software Help Sessions in Three Minutes or Less.” In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–13. ISBN: 9781450367080. DOI: [10.1145/3313831.3376230](https://doi.org/10.1145/3313831.3376230).
- [85] Daniel Kahneman. *Thinking, Fast and Slow*. 2011.
- [86] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. “ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education.” In: *Learning and Individual Differences* 103 (2023), p. 102274. ISSN: 1041-6080. DOI: <https://doi.org/10.1016/j.lindif.2023.102274>.
- [87] Ronald T. Kellogg. “Training writing skills: A cognitive developmental perspective.” In: *Journal of Writing Research* 1.1 (2008), pp. 1–26. DOI: [10.17239/jowr-2008.01.01.1](https://doi.org/10.17239/jowr-2008.01.01.1).
- [88] Sebastian Kernbach and Sabrina Bresciani. “10 Years after Tufte’s “Cognitive Style of PowerPoint”: Synthesizing its Constraining Qualities.” In: *2013 17th International Conference on Information Visualisation*. 2013, pp. 345–350. DOI: [10.1109/IV.2013.44](https://doi.org/10.1109/IV.2013.44).
- [89] Sebastian Kernbach, Sabrina Bresciani, and Martin J. Eppler. “Slip-Sliding-Away: A Review of the Literature on the Constraining Qualities of PowerPoint.” In: *Business and Professional Communication Quarterly* 78.3 (2015), pp. 292–313. DOI: [10.1177/2329490615595499](https://doi.org/10.1177/2329490615595499).
- [90] Kenneth A. Kiewra. “Investigating Notetaking and Review: A Depth of Processing Alternative.” In: *Educational Psychologist* 20.1 (1985), pp. 23–32. DOI: [10.1207/s15326985ep2001_4](https://doi.org/10.1207/s15326985ep2001_4).

- [91] Kenneth A. Kiewra. "A Review of Note-Taking: The Encoding-Storage Paradigm and Beyond." In: *Educational Psychology Review* 1.2 (1989), pp. 147–172. DOI: [10.1007/BF01326640](https://doi.org/10.1007/BF01326640).
- [92] Daekun Kim, Nikhita Joshi, and Daniel Vogel. "Perspective and Geometry Approaches to Mouse Cursor Control in Spatial Augmented Reality." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: [10.1145/3544548.3580849](https://doi.org/10.1145/3544548.3580849).
- [93] Jaejeung Kim, Hayoung Jung, Minsam Ko, and Uichin Lee. "Goal-Keeper: Exploring Interaction Lockout Mechanisms for Regulating Smartphone Use." In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3.1 (2019). DOI: [10.1145/3314403](https://doi.org/10.1145/3314403).
- [94] Jaejeung Kim, Joonyoung Park, Hyunsoo Lee, Minsam Ko, and Uichin Lee. "LocknType: Lockout Task Intervention for Discouraging Smartphone App Use." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–12. ISBN: 9781450359702. DOI: [10.1145/3290605.3300927](https://doi.org/10.1145/3290605.3300927).
- [95] Jaewon Kim, Paul Thomas, Ramesh Sankaranarayana, Tom Gedeon, and Hwan-Jin Yoon. "Pagination versus Scrolling in Mobile Web Search." In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. CIKM '16. Indianapolis, Indiana, USA: Association for Computing Machinery, 2016, 751–760. ISBN: 9781450340731. DOI: [10.1145/2983323.2983720](https://doi.org/10.1145/2983323.2983720).
- [96] Keiichi Kobayashi. "What limits the encoding effect of note-taking? A meta-analytic examination." In: *Contemporary Educational Psychology* 30.2 (2005), pp. 242–262. ISSN: 0361-476X. DOI: [10.1016/j.cedpsych.2004.10.001](https://doi.org/10.1016/j.cedpsych.2004.10.001).
- [97] Stefan Köpsell. "Low Latency Anonymous Communication – How Long Are Users Willing to Wait?" In: *Emerging Trends in Information and Communication Security*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 221–237. ISBN: 978-3-540-34642-5. DOI: [10.1007/11766155_16](https://doi.org/10.1007/11766155_16).
- [98] Pia Kreijkes, Viktor Kewenig, Martina Kuvalja, Mina Lee, Sylvia Vitello, Jake M. Hofman, Abigail Sellen, Sean Rintel, Daniel G. Goldstein, David M. Rothschild, Lev Tankelevitch, and Tim Oates. *Effects of LLM Use and Note-Taking on Reading Comprehension and Memory: A Randomised Experiment in Secondary Schools*. <http://dx.doi.org/10.2139/ssrn.5095149>. 2025.
- [99] Sneha R. Krishna Kumaran, Yue Yin, and Brian P. Bailey. "Plan Early, Revise More: Effects of Goal Setting and Perceived Role of the Feedback Provider on Feedback Seeking Behavior." In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: [10.1145/3449098](https://doi.org/10.1145/3449098).

- [100] Ida Larsen-Ledet and Marcel Borowski. ““It Looks Like You Don’t Agree”: Idiosyncratic Practices and Preferences in Collaborative Writing.” In: *32nd Australian Conference on Human-Computer Interaction. OzCHI ’20*. Sydney, NSW, Australia: Association for Computing Machinery, 2020, 339–354. ISBN: 9781450389754. DOI: [10.1145/3441000.3441032](https://doi.org/10.1145/3441000.3441032).
- [101] Ida Larsen-Ledet and Henrik Korsgaard. “Territorial Functioning in Collaborative Writing: Fragmented Exchanges and Common Outcomes.” In: *Comput. Supported Coop. Work* 28.3–4 (2019), 391–433. ISSN: 0925-9724. DOI: [10.1007/s10606-019-09359-8](https://doi.org/10.1007/s10606-019-09359-8).
- [102] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Tiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Serman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. “A Design Space for Intelligent and Interactive Writing Assistants.” In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: [10.1145/3613904.3642697](https://doi.org/10.1145/3613904.3642697).
- [103] Mina Lee, Percy Liang, and Qian Yang. “CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities.” In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: [10.1145/3491102.3502030](https://doi.org/10.1145/3491102.3502030).
- [104] Florian Lehmann, Niklas Markert, Hai Dang, and Daniel Buschek. “Suggestion Lists vs. Continuous Generation: Interaction Design for Writing with Generative Models on Mobile Devices Affect Text Length, Wording and Perceived Authorship.” In: *Proceedings of Mensch Und Computer 2022*. MuC ’22. Darmstadt, Germany: Association for Computing Machinery, 2022, 192–208. ISBN: 9781450396905. DOI: [10.1145/3543758.3543947](https://doi.org/10.1145/3543758.3543947).
- [105] Detlev Leutner, Claudia Leopold, and Viola den Elzen-Rump. “Self-Regulated Learning with a Text-Highlighting Strategy.” In: *Zeitschrift für Psychologie/Journal of Psychology* 215.3 (2007), pp. 174–182. DOI: [10.1027/0044-3409.215.3.174](https://doi.org/10.1027/0044-3409.215.3.174).
- [106] Blaine Lewis and Daniel Vogel. “Longer Delays in Rehearsal-based Interfaces Increase Expert Use.” In: *ACM Trans. Comput.-Hum. Interact.* 27.6 (2020). ISSN: 1073-0516. DOI: [10.1145/3418196](https://doi.org/10.1145/3418196).
- [107] Blaine Lewis, Greg d’Eon, Andy Cockburn, and Daniel Vogel. “KeyMap: Improving Keyboard Shortcut Vocabulary Using Norman’s Mapping.” In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for

- Computing Machinery, 2020, 1–10. ISBN: 9781450367080. DOI: [10.1145/3313831.3376483](https://doi.org/10.1145/3313831.3376483).
- [108] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. “The Value, Benefits, and Concerns of Generative AI-Powered Assistance in Writing.” In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: [10.1145/3613904.3642625](https://doi.org/10.1145/3613904.3642625).
 - [109] Lassi A. Liikkanen, Tua A. Björklund, Matti M. Hämäläinen, and Mikko P. Koskinen. “Time Constraints in Design Idea Generation.” In: *DS 58-9: Proceedings of ICED 09, the 17th International Conference on Engineering Design, Vol. 9, Human Behavior in Design, Palo Alto, CA, USA, 24.-27.08. 2009*. 2009, pp. 81–90.
 - [110] Ming Liu, Jingxu Zhang, Lucy Michael Nyagoga, and Li Liu. “Student-AI Question Cocreation for Enhancing Reading Comprehension.” In: *IEEE Transactions on Learning Technologies* 17 (2024), pp. 815–826. DOI: [10.1109/TLT.2023.3333439](https://doi.org/10.1109/TLT.2023.3333439).
 - [111] Kirsti Lonka, Sari Lindblom-YlÄänne, and Sini Maury. “The Effect of Study Strategies on Learning from Text.” In: *Learning and Instruction* 4.3 (1994), pp. 253–271. ISSN: 0959-4752. DOI: [10.1016/0959-4752\(94\)90026-4](https://doi.org/10.1016/0959-4752(94)90026-4).
 - [112] Xing Lu and Zhicong Lu. “Fifteen Seconds of Fame: A Qualitative Study of Douyin, A Short Video Sharing Mobile Application in China.” In: *Social Computing and Social Media. Design, Human Behavior and Analytics: 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part I*. Orlando, FL, USA: Springer-Verlag, 2019, 233–244. ISBN: 978-3-030-21901-7. DOI: [10.1007/978-3-030-21902-4_17](https://doi.org/10.1007/978-3-030-21902-4_17).
 - [113] Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. “Structuring, Aggregating, and Evaluating Crowdsourced Design Critique.” In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW ’15. Vancouver, BC, Canada: Association for Computing Machinery, 2015, 473–485. ISBN: 9781450329224. DOI: [10.1145/2675133.2675283](https://doi.org/10.1145/2675133.2675283).
 - [114] Thor Magnusson. “Designing Constraints: Composing and Performing with Digital Musical Systems.” In: *Computer Music Journal* 34.4 (2010), pp. 62–73.
 - [115] Catherine C. Marshall. “Annotation: from paper books to the digital library.” In: *Proceedings of the Second ACM International Conference on Digital Libraries*. DL ’97. Philadelphia, Pennsylvania, USA: Association for Computing Machinery, 1997, 131–140. ISBN: 0897918681. DOI: [10.1145/263690.263806](https://doi.org/10.1145/263690.263806).

- [116] Catherine C. Marshall and A. J. Bernheim Brush. "Exploring the relationship between personal and public annotations." In: *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*. JCDL '04. Tucson, AZ, USA: Association for Computing Machinery, 2004, 349–357. ISBN: 1581138326. DOI: [10.1145/996350.996432](https://doi.org/10.1145/996350.996432).
- [117] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. "Charagraph: Interactive Generation of Charts for Realtime Annotation of Data-Rich Paragraphs." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: [10.1145/3544548.3581091](https://doi.org/10.1145/3544548.3581091).
- [118] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. "DirectGPT: A Direct Manipulation Interface to Interact with Large Language Models." In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: [10.1145/3613904.3642462](https://doi.org/10.1145/3613904.3642462).
- [119] Liam Melin-Higgins. *AI in Document Interaction: An Interaction Design Approach to Enhancing Reading Practices*. 2024.
- [120] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. "Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: [10.1145/3544548.3581225](https://doi.org/10.1145/3544548.3581225).
- [121] Adam Morgan and Mark Barden. *A Beautiful Constraint: How to Transform Your Limitations Into Advantages, and Why It's Everyone's Business*. John Wiley & Sons, 2015.
- [122] Fiona Fui-Hoon Nah. "A Study on Tolerable Waiting Time: How Long are Web Users Willing to Wait?" In: *Behaviour & Information Technology* 23.3 (2004), pp. 153–163. DOI: [10.1080/01449290410001669914](https://doi.org/10.1080/01449290410001669914).
- [123] Timothy Neate, Abi Roper, Stephanie Wilson, and Jane Marshall. "Empowering Expression for Users with Aphasia through Constrained Creativity." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–12. ISBN: 9781450359702. DOI: [10.1145/3290605.3300615](https://doi.org/10.1145/3290605.3300615).
- [124] Justin Nicholes. *Measuring Ownership of Creative Versus Academic Writing: Implications for Interdisciplinary Praxis*. 2017. DOI: [10.62959/WIP-03-2017-08](https://doi.org/10.62959/WIP-03-2017-08).
- [125] Jakob Nielsen. *Response Times: The 3 Important Limits*. <https://www.nngroup.com/articles/response-times-3-important-limits/>. [Accessed 07-10-2024]. 1993.
- [126] Jakob Nielsen. *Legibility, Readability, and Comprehension: Making Users Read Your Words*. <https://www.nngroup.com/articles/legibility-readability-comprehension/>. [Accessed 08-12-2024]. 2015.

- [127] Sherrie L. Nist and Mark C. Hogrebe. "The Role of Underlining and Annotating in Remembering Textual Information." In: *Reading Research and Instruction* 27.1 (1987), pp. 12–25. DOI: [10.1080/19388078709557922](https://doi.org/10.1080/19388078709557922).
- [128] Don Norman. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books, 2013.
- [129] Donald A. Norman. "Affordance, Conventions, and Design." In: *Interactions* 6.3 (May 1999), 38–43. ISSN: 1072-5520. DOI: [10.1145/301153.301168](https://doi.org/10.1145/301153.301168).
- [130] NoteGPT. *NoteGPT – Your All-in-One AI Learning Assistant. Summarize, Chat & Write – Fast & Free*. <https://notegpt.io>. [Accessed 20-08-2025]. 2025.
- [131] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. "I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence." In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal, QC, Canada: Association for Computing Machinery, 2018, 1–13. ISBN: 9781450356206. DOI: [10.1145/3173574.3174223](https://doi.org/10.1145/3173574.3174223).
- [132] Thierry Olive and Marie-Laure Barbier. "Processing Time and Cognitive Effort of Longhand Note Taking When Reading and Summarizing a Structured or Linear Text." In: *Written Communication* 34.2 (2017), pp. 224–246. DOI: [10.1177/0741088317699898](https://doi.org/10.1177/0741088317699898).
- [133] Judith S. Olson, Dakuo Wang, Gary M. Olson, and Jingwen Zhang. "How People Write Together Now: Beginning the Investigation with Advanced Undergraduates in a Project Course." In: *ACM Trans. Comput.-Hum. Interact.* 24.1 (2017). ISSN: 1073-0516. DOI: [10.1145/3038919](https://doi.org/10.1145/3038919).
- [134] Balder Onarheim and Michael Mose Biskjaer. "Balancing Constraints and the Sweet Spot as Coming Topics for Creativity Research." In: *Creativity in Design: Understanding, Capturing, Supporting* 1 (2015), pp. 1–18.
- [135] Antti Oulasvirta, Janne P. Hukkinen, and Barry Schwartz. "When More is Less: The Paradox of Choice in Search Engine Use." In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '09. Boston, MA, USA: Association for Computing Machinery, 2009, 516–523. ISBN: 9781605584836. DOI: [10.1145/1571941.1572030](https://doi.org/10.1145/1571941.1572030).
- [136] Marcus Pearce and Geraint A. Wiggins. "Aspects of a Cognitive Theory of Creativity in Musical Composition." In: *Proceedings of the ECAIo2 Workshop on Creative Systems*. 2002.
- [137] James Pierce and Eric Paulos. "Counterfunctional Things: Exploring Possibilities in Designing Digital Limitations." In: *Proceedings of the 2014 Conference on Designing Interactive Systems*. DIS '14. Vancouver, BC, Canada: Association for Computing Machinery, 2014, 375–384. ISBN: 9781450329026. DOI: [10.1145/2598510.2598522](https://doi.org/10.1145/2598510.2598522).

- [138] Jon L. Pierce, Tatiana Kostova, and Kurt T. Dirks. "Toward a Theory of Psychological Ownership in Organizations." In: *The Academy of Management Review* 26.2 (2001), pp. 298–310. ISSN: 03637425. DOI: [10.2307/259124](https://doi.org/10.2307/259124).
- [139] Jon L. Pierce, Tatiana Kostova, and Kurt T. Dirks. "The State of Psychological Ownership: Integrating and Extending a Century of Research." In: *Review of General Psychology* 7.1 (2003), pp. 84–107. DOI: [10.1037/1089-2680.7.1.84](https://doi.org/10.1037/1089-2680.7.1.84).
- [140] Annie Piolat, Thierry Olive, and Ronald T Kellogg. "Cognitive Effort during Note Taking." In: *Applied Cognitive Psychology* 19.3 (2005), pp. 291–312. DOI: [10.1002/acp.1086](https://doi.org/10.1002/acp.1086).
- [141] Annie Piolat, Jean-Yves Roussey, and Olivier Thunin. "Effects of screen presentation on text reading and revising." In: *International Journal of Human-Computer Studies* 47.4 (1997), pp. 565–589. DOI: [10.1006/ijhc.1997.0145](https://doi.org/10.1006/ijhc.1997.0145).
- [142] Peter Pirolli and Stuart Card. "The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis." In: *Proceedings of International Conference on Intelligence Analysis*. Vol. 5. McLean, VA, USA. 2005, pp. 2–4.
- [143] Carol Porter-O'Donnell. "Beyond the Yellow Highlighter: Teaching Annotation Skills to Improve Reading Comprehension." In: *English Journal* 93.Secondary Readers Reading Successfully (2004), pp. 82–89. DOI: [10.58680/ej20044802](https://doi.org/10.58680/ej20044802).
- [144] Ilona R. Posner and Ronald M. Baecker. "How People Write Together (Groupware)." In: *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*. Vol. iv. 1992, 127–138 vol.4. DOI: [10.1109/HICSS.1992.183420](https://doi.org/10.1109/HICSS.1992.183420).
- [145] Melanie Ramdarshan Bold and Kiri L. Wagstaff. "Marginalia in the digital age: Are digital reading devices meeting the needs of today's readers?" In: *Library & Information Science Research* 39.1 (2017), pp. 16–22. ISSN: 0740-8188. DOI: [10.1016/j.lisr.2017.01.004](https://doi.org/10.1016/j.lisr.2017.01.004).
- [146] Walter R. Reitman. "Heuristic decision procedures, open constraints, and the structure of ill-defined problems." In: *Human Judgments and Optimality* (1964), pp. 282–315.
- [147] Miguel A. Renom, Baptiste Caramiaux, and Michel Beaudouin-Lafon. "Exploring Technical Reasoning in Digital Tool Use." In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: [10.1145/3491102.3501877](https://doi.org/10.1145/3491102.3501877).
- [148] Jeba Rezwana and Mary Lou Maher. "User Perspectives on Ethical Challenges in Human-AI Co-Creativity: A Design Fiction Study." In: *Proceedings of the 15th Conference on Creativity and Cognition*. C&C '23. Virtual Event, USA: Association for Computing Machinery, 2023, 62–74. ISBN: 9798400701801. DOI: [10.1145/3591196.3593364](https://doi.org/10.1145/3591196.3593364).

- [149] Grégoire Richard, Thomas Pietrzak, Ferran Argelaguet, Anatole Lécuyer, and Géry Casiez. "Within or Between? Comparing Experimental Designs for Virtual Embodiment Studies." In: *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 2022, pp. 186–195. DOI: [10.1109/VR51125.2022.00037](https://doi.org/10.1109/VR51125.2022.00037).
- [150] Hugo Romat, Emmanuel Pietriga, Nathalie Henry-Riche, Ken Hinckley, and Caroline Appert. "SpaceInk: Making Space for In-Context Annotations." In: *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. UIST '19. New Orleans, LA, USA: Association for Computing Machinery, 2019, 871–882. ISBN: 9781450368162. DOI: [10.1145/3332165.3347934](https://doi.org/10.1145/3332165.3347934).
- [151] Brent D. Rosso. "Creativity and Constraints: Exploring the Role of Constraints in the Creative Processes of Research and Development Teams." In: *Organization Studies* 35.4 (2014), pp. 551–585. DOI: [10.1177/017084061351760](https://doi.org/10.1177/017084061351760).
- [152] Quentin Roy, Futian Zhang, and Daniel Vogel. "Automation Accuracy Is Good, but High Controllability May Be Better." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–8. ISBN: 9781450359702. DOI: [10.1145/3290605.3300750](https://doi.org/10.1145/3290605.3300750).
- [153] Timothy Ryan. *Fraudulent responses on Amazon Mechanical Turk: A Fresh Cautionary Tale*. [Accessed 09-04-2022]. 2020. URL: <https://timryan.web.unc.edu/2020/12/22/fraudulent-responses-on-amazon-mechanical-turk-a-fresh-cautionary-tale/>.
- [154] Christopher A. Sanchez and Jennifer Wiley. "To Scroll or Not to Scroll: Scrolling, Working Memory Capacity, and Comprehending Complex Texts." In: *Human Factors* 51.5 (2009), pp. 730–738. DOI: [10.1177/0018720809352788](https://doi.org/10.1177/0018720809352788).
- [155] Paula Selvidge. "How Long is Too Long to Wait for a Website to Load." In: *Usability News* 1.2 (1999), pp. 1–3.
- [156] Selina Sharmin, Oleg Špakov, and Kari-Jouko Räihä. "Reading on-screen text with gaze-based auto-scrolling." In: *Proceedings of the 2013 Conference on Eye Tracking South Africa*. ETSA '13. Cape Town, South Africa: Association for Computing Machinery, 2013, 24–31. ISBN: 9781450321105. DOI: [10.1145/2509315.2509319](https://doi.org/10.1145/2509315.2509319).
- [157] Richard Shay, Saranga Komanduri, Adam L. Durity, Phillip (Seyoung) Huh, Michelle L. Mazurek, Sean M. Segreti, Blase Ur, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. "Can long passwords be secure and usable?" In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. Toronto, Ontario, Canada: Association for Computing Machinery, 2014, 2927–2936. ISBN: 9781450324731. DOI: [10.1145/2556288.2557377](https://doi.org/10.1145/2556288.2557377).
- [158] Hua Shen and Tongshuang Wu. *Parachute: Evaluating Interactive Human-LM Co-writing Systems*. 2023. arXiv: [2303.06333](https://arxiv.org/abs/2303.06333) [cs.HC].

- [159] Ben Shneiderman. "Response Time and Display Rate in Human Performance with Computers." In: *ACM Comput. Surv.* 16.3 (Sept. 1984), 265–285. ISSN: 0360-0300. DOI: [10.1145/2514.2517](https://doi.org/10.1145/2514.2517).
- [160] Pao Siangliulue, Joel Chan, Krzysztof Z. Gajos, and Steven P. Dow. "Providing Timely Examples Improves the Quantity and Quality of Generated Ideas." In: *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. C&C '15. Glasgow, United Kingdom: Association for Computing Machinery, 2015, 83–92. ISBN: 9781450335980. DOI: [10.1145/2757226.2757230](https://doi.org/10.1145/2757226.2757230).
- [161] Herbert A. Simon. "The Structure of Ill Structured Problems." In: *Artificial Intelligence* 4.3 (1973), pp. 181–201. ISSN: 0004-3702. DOI: [https://doi.org/10.1016/0004-3702\(73\)90011-8](https://doi.org/10.1016/0004-3702(73)90011-8).
- [162] Simpleview. [Accessed 09-07-2023]. 2023. URL: <https://www.simpleviewinc.com/blog/stories/post/adding-read-time-on-blogs-boosts-engagement-by-up-to-40/>.
- [163] Ana Cristina O. Siqueira and Benson Honig. "Entrepreneurs' ingenuity and self-imposed ethical constraints: creating sustainability-oriented new ventures and knowledge." In: *Journal of Knowledge Management* 23.10 (Sept. 2019), pp. 1965–1983. ISSN: 1367-3270. DOI: [10.1108/JKM-11-2018-0707](https://doi.org/10.1108/JKM-11-2018-0707).
- [164] Frank Soboczenski, Paul Cairns, and Anna L. Cox. "Increasing Accuracy by Decreasing Presentation Quality in Transcription Tasks." In: *Human-Computer Interaction—INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part II* 14. Springer. 2013, pp. 380–394. DOI: [10.1007/978-3-642-40480-1_25](https://doi.org/10.1007/978-3-642-40480-1_25).
- [165] Elizabeth G. Soslau and Deborah S. Yost. "Urban Service-Learning: An Authentic Teaching Strategy to Deliver a Standards-Driven Curriculum." In: *Journal of Experiential Education* 30.1 (2007), pp. 36–53. DOI: [10.1177/105382590703000104](https://doi.org/10.1177/105382590703000104).
- [166] Martin Stacey and Claudia Eckert. "Reshaping the box: creative designing as constraint management." In: *International Journal of Product Development* 11.3-4 (2010), pp. 241–255. DOI: [10.1504/IJPD.2010.03396](https://doi.org/10.1504/IJPD.2010.03396).
- [167] Patricia D. Stokes. *Creativity from Constraints: The Psychology of Breakthrough*. Springer Publishing Company, 2005.
- [168] Carla Tamburro, Timothy Neate, Abi Roper, and Stephanie Wilson. "Accessible Creativity with a Comic Spin." In: *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS '20. Virtual Event, Greece: Association for Computing Machinery, 2020. ISBN: 9781450371032. DOI: [10.1145/3373625.3417012](https://doi.org/10.1145/3373625.3417012).
- [169] Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT, USA: Yale University Press, 2008. ISBN: 9780300146813.

- [170] Jakob Tholander and Maria Normark. "Crafting Personal Information - Resistance, Imperfection, and Self-Creation in Bullet Journaling." In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–13. ISBN: 9781450367080. DOI: [10.1145/3313831.3376410](https://doi.org/10.1145/3313831.3376410).
- [171] Jennifer Thom-Santelli, Dan R. Cosley, and Geri Gay. "What's Mine is Mine: Territoriality in Collaborative Authoring." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. Boston, MA, USA: Association for Computing Machinery, 2009, 1481–1484. ISBN: 9781605582467. DOI: [10.1145/1518701.1518925](https://doi.org/10.1145/1518701.1518925).
- [172] Theophanis Tsandilas and Géry Casiez. *The illusory promise of the aligned rank transform*. <https://statransform.github.io/jovi/>. 2025.
- [173] Sherman W., Paula T. Hertel, Marvin C. McCallum, and Henry C. Ellis. "Cognitive Effort and Memory." In: *Journal of Experimental Psychology: Human Learning and Memory* 5.6 (1979), p. 607. DOI: [10.1037/0278-7393.5.6.607](https://doi.org/10.1037/0278-7393.5.6.607).
- [174] Jeffrey J. Walczyk, Kathryn E. Kelly, Scott D. Meche, and Hillary Braud. "Time Limitations Enhance Reading Comprehension." In: *Contemporary educational psychology* 24.2 (1999), pp. 156–165. DOI: [10.1006/ceps.1998.0992](https://doi.org/10.1006/ceps.1998.0992).
- [175] Shaun Wallace, Zoya Bylinskii, Jonathan Dobres, Bernard Kerr, Sam Berlow, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Dave B. Miller, Jeff Huang, and Ben D. Sawyer. "Towards Individuated Reading Experiences: Different Fonts Increase Reading Speed for Different Individuals." In: *ACM Trans. Comput.-Hum. Interact.* 29.4 (Mar. 2022). ISSN: 1073-0516. DOI: [10.1145/3502222](https://doi.org/10.1145/3502222).
- [176] William P. Wallace. "Review of the historical, empirical, and theoretical status of the von Restorff phenomenon." In: *Psychological Bulletin* 63.6 (1965), pp. 410–424. DOI: [10.1037/h0022001](https://doi.org/10.1037/h0022001).
- [177] Jin Wang and Wenxiang Fan. "The Effect of ChatGPT on Students' Learning Performance, Learning Perception, and Higher-Order Thinking: Insights from a Meta-Analysis." In: *Humanities and Social Sciences Communications* 12.1 (2025), pp. 1–21. DOI: [10.1057/s41599-025-04787-y](https://doi.org/10.1057/s41599-025-04787-y).
- [178] QianYing Wang, Alberto Battocchi, Ilenia Graziola, Fabio Pianesi, Daniel Tomasini, Massimo Zancanaro, and Clifford Nass. "The role of psychological ownership and ownership markers in collaborative working environment." In: *Proceedings of the 8th International Conference on Multimodal Interfaces*. ICMI '06. Banff, Alberta, Canada: Association for Computing Machinery, 2006, 225–232. ISBN: 159593541X. DOI: [10.1145/1180995.1181041](https://doi.org/10.1145/1180995.1181041).
- [179] Xizhe Wang, Yihua Zhong, Changqin Huang, and Xiaodi Huang. "ChatPRCS: A Personalized Support System for English Reading Comprehension Based on ChatGPT." In: *IEEE Transactions on Learning Technologies* 17 (2024), pp. 1722–1736. DOI: [10.1109/TLT.2024.3405747](https://doi.org/10.1109/TLT.2024.3405747).

- [180] Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. "A Field Trial of Privacy Nudges for Facebook." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. Toronto, Ontario, Canada: Association for Computing Machinery, 2014, 2367–2376. ISBN: 9781450324731. DOI: [10.1145/2556288.2557413](https://doi.org/10.1145/2556288.2557413).
- [181] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. "The Aligned Rank Transform for Nonparametric Factorial Analyses using only Anova Procedures." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada: Association for Computing Machinery, 2011, 143–146. ISBN: 9781450302289. DOI: [10.1145/1978942.1978963](https://doi.org/10.1145/1978942.1978963).
- [182] Jacob O. Wobbrock and Julie A. Kientz. "Research Contributions in Human-Computer Interaction." In: *Interactions* 23.3 (Apr. 2016), 38–44. ISSN: 1072-5520. DOI: [10.1145/2907069](https://doi.org/10.1145/2907069).
- [183] Maryanne Wolf, Mirit Barzillai, and John Dunne. "The Importance of Deep Reading." In: *Challenging the Whole Child: Reflections on Best Practices in Learning, Teaching, and Leadership* 130.21 (2009).
- [184] Joanna L. Wolfe and Christine M. Neuwirth. "From the Margins to the Center: The Future of Annotation." In: *Journal of Business and Technical Communication* 15.3 (2001), pp. 333–371. DOI: [10.1177/105065190101500304](https://doi.org/10.1177/105065190101500304).
- [185] Erik Wästlund, Torsten Norlander, and Trevor Archer. "The effect of page layout on mental workload: A dual-task experiment." In: *Computers in Human Behavior* 24.3 (2008), pp. 1229–1245. ISSN: 0747-5632. DOI: [10.1016/j.chb.2007.05.001](https://doi.org/10.1016/j.chb.2007.05.001).
- [186] Catherine Yeh, Gonzalo Ramos, Rachel Ng, Andy Huntington, and Richard Banks. *GhostWriter: Augmenting Collaborative Human-AI Writing Experiences Through Personalization and Agency*. 2025. arXiv: [2402.08855](https://arxiv.org/abs/2402.08855) [cs.HC].
- [187] Yen-Ting Yeh, Nikhita Joshi, and Daniel Vogel. "The Effects of Update Interval and Reveal Method on Writer Comfort in Synchronized Shared-Editors." In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Honolulu, HI, USA: Association for Computing Machinery, 2024, pp. 1–13. DOI: [10.1145/3613904.3642330](https://doi.org/10.1145/3613904.3642330).
- [188] Ryan Yen, Jian Zhao, and Daniel Vogel. "Code Shaping: Iterative Code Editing with Free-form AI-Interpreted Sketching." In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI '25. New York, NY, USA: Association for Computing Machinery, 2025. ISBN: 9798400713941. DOI: [10.1145/3706598.3713822](https://doi.org/10.1145/3706598.3713822).
- [189] Soobin Yim, Dakuo Wang, Judith Olson, Viet Vu, and Mark Warschauer. "Synchronous Collaborative Writing in the Classroom: Undergraduates' Collaboration Practices and their Impact on Writing Style, Quality, and Quantity." In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. Portland, Oregon, USA: Association for Computing Machinery, 2017, 468–479. ISBN: 9781450343350. DOI: [10.1145/2998181.2998356](https://doi.org/10.1145/2998181.2998356).

- [190] H. Peyton Young. "Condorcet's Theory of Voting." In: *American Political Science Review* 82.4 (1988), pp. 1231–1244.
- [191] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. "Wordcraft: Story Writing With Large Language Models." In: *27th International Conference on Intelligent User Interfaces*. IUI '22. Helsinki, Finland: Association for Computing Machinery, 2022, 841–852. ISBN: 9781450391443. DOI: [10.1145/3490099.3511105](https://doi.org/10.1145/3490099.3511105).
- [192] Carole L. Yue, Benjamin C. Storm, Nate Kornell, and Elizabeth Ligon Bjork. "Highlighting and Its Relation to Distributed Study and Students' Metacognitive Beliefs." In: *Educational Psychology Review* 27.1 (2015), pp. 69–78. DOI: [10.1007/s10648-014-9277-z](https://doi.org/10.1007/s10648-014-9277-z).

GLOSSARY

active reading reading a document and engaging with the text, for example, through annotations. [3](#), [23](#), [24](#), [82](#), [109](#)

dual process theory a theory claiming that humans have two ways of thinking: fast and automatic “System 1” thinking, and slow and deliberate “System 2” thinking. [10](#), [42](#), [110](#)

levels of processing theory a theory claiming that humans can remember information better the deeper it has been processed. [24](#), [84](#)

nudge encouraging a specific outcome without preventing others. [10](#), [42](#), [66](#)

passive reading reading a document without engaging with the text at a deeper level. [3](#), [13](#), [109](#)

psychological ownership feelings of something belonging to someone, regardless of legal ownership. [3](#), [44](#), [83](#), [87](#)

working memory capacity a cognitive system describing the amount of information humans can process and store simultaneously. [14](#), [87](#)