Breaking the Working Memory Barrier: Introducing a Novel Digital Tool to Improve Understanding of Mathematical Texts

Breaking the Working Memory Barrier: Introducing a Novel Digital Tool to Improve Understanding of Mathematical Texts

A THESIS PRESENTED
BY
LINDA QIN
TO
THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS
IN THE SUBJECT OF
COMPUTER SCIENCE (MIND, BRAIN AND BEHAVIOR)

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2020

Thesis advisor: Professor Jim Waldo and Krzysztof Gajos Linda Qin

Breaking the Working Memory Barrier: Introducing a Novel Digital Tool to Improve Understanding of Mathematical Texts

Abstract

With the digitalization of academic material, there exists opportunity to utilize external tools to improve understanding of notation-heavy text. Prior research has shown a connection between working memory limits and conceptual learning. This study examines the creation and efficacy of a digital tool that acts as a crutch for working memory, through saving and resurfacing variable definitions across texts. 46 Harvard students in concentrating in STEM and Economics were asked to read two articles by the same author, one using the tool and one without, before answering questions about the excerpts about main arguments, detailed equations, and connections to related topics. These responses were then graded on a 5-point scale using Amazon Mechanical Turk. Significant results were found in the recall of specific equations as well as participant confidence in understanding the text, especially for participants with no experience reading similar-caliber text, but insignificance in grasping larger conceptual ideas. Further research is suggested in sample size and experimental design in order to analyze the longer-term effects of this tool on learning and memory.

Contents

This thesis is dedicated to women in STEM, especially those with the courage to pursue these fields for the first time in college.

Acknowledgments

It takes a village. I would not have been able to do this without the help of many along the way, starting from my thesis advisors, Jim Waldo and Krzysztof Gajos, for your infinite patience and wisdom. Thank you to all those who have supported me through the long nights, including but not limited to my blockmates Ching Sullivan, Karen Yang, Angel Zhou, Fay Huo, Sierra Tseng, and Myung-seo Yoon, my thesis fairy Sabrina Chok, Kung Fu Tea, and Shin ramen. I extend eternal gratitude for all those who participated in my study, both on-campus and remotely, especially for those who came during our last week on campus whilst simultaneously struggling through moving-out, saying goodbye, and ensuring the safety of your loved ones. Finally, to mom and dad: gratias maximas ago. Thank you for everything you gave me during this journey, both in my thesis and at Harvard. I hope this makes you proud.

1

Introduction

There exists a desire to improve the efficiency of reading and understanding notation-heavy text, whether it be students with course textbooks or academics with scholarly journals. In many cases, it takes multiple reads of the text or a significantly slower reading pace in order to fully grasp the ideas the authors propose. However, as technology becomes increasingly utilized in academia, the digitalization of texts creates an opportunity to augment the reading experience in a way that was previously not possible with physical texts. How can we take advantage of this new digital era to decrease reading time while maintaining or even improving understanding of the material?

The structure of academic writings can be broken down into three parts: the introduction of the concept, its presentation, and the analysis of the aforementioned concept. In cases where mathematical notation is involved, variables are usually defined during the initial description of the idea. However, these introductory sections can be quite lengthy, resulting in the reader needing to keep track of many moving pieces before even reaching the equations themselves, especially if many different variables are introduced. This results in excess time spent parsing through and referencing previously-defined variables before the reader is able to grasp the meaning of the equations presented.

I now introduce the concept of working memory, a type of short-term memory specifically set aside for cognitive tasks. [2] I hypothesize that it is due to the limitations of working memory that a non-neglible amount of time is spent parsing through variables in notation-heavy documents, as there is not enough memory capacity to temporarily store all the variables' descriptions. This results in the need to context-switch and cross-reference previously-defined variables with the given equations, which is a very inefficient process to internalize their definitions. By streamlining this process and cutting off excess time, I hypothesize that readers are able to reach both a faster and deeper understanding of notation-heavy material. Therefore, in this paper, I introduce a tool that enables readers to keep track of variables as they are presented, acting as a crutch to the limitations of working memory.

1.1 Related Work

While a similar tool does not yet exist, research has been conducted on the use of working memory in cognitively-intense settings. Multiple studies (Danili & Reid, 2004[3], Peng et al., 2015[6]) have established a correlation between working memory and problem solving, the latter focusing especially in the field of mathematics. The former examined grade 10 students in Greece's performance across three test types, multiple choice, short answer, and a structural communication grid, in the context of psychological information, and discovered that working memory specifically plays a significant role in determining the rate of conceptual learning.

Cowan (2008) creates a theoretical framework for working memory and provides a great example to highlight the importance of memory in learning, paraphrased in the following sentences. To understand the transitive property in algebra, that, since a + b = c + d and c + d = e, a + b = e, one must not only be able to recall algebraic rules but also maintain both equations in memory. Specifically, even if the equations are placed adjacently on the page, this conclusion may not be reached unless both are concurrently encoded into working memory. As the entities and concepts become more complex, more working memory is required to reach the appropriate conclusions.

Since working memory is so closely tied with learning, many studies in this area have focused on global student experiences with STEM subjects. Hindal et al. (2009)[4] found a cause-effect relationship between Saudi Arabian students' working memory capacity and their abilities in science, while Jung and Reid (2009)[5] further concluded through similar examination-based studies that the memory-performance connection affected South

Korean students' attitudes toward STEM studies. As working memory is highly heritable, students have a level of genetic predisposition toward STEM fields. Those with a high working memory capacity focus more on grasping concepts and theories, which proves more beneficial in the long run, compared to those with a lower capacity who tend to resort more on rote memorization. This finding is especially relevant for the college student demographic, as performance in high school and early college greatly affect attitudes toward future fields of study.

Furthermore, while working memory is a large focus of this study, the following experiment also analyzes users' ability to comprehend mathematical text. Acar [1] analyzed the readability and comprehensibility of technical and scientific texts with the basis that comprehensibility relies on a mix of inclusion of straightforward, technical elements and simplicity of conveyance of message. The experimental portion of this thesis borrows this idea and analyzes understanding of excerpts read in the context of both specific technical details and overall arguments.

While extensive research has been conducted on the connection between working memory and mathematics, there exists very few documented works tackling some of the aforementioned issues (outside of recommendations for specific education and teaching methods). This presents an opportunity to explore additional methods to democratize the reading of cognitively-heavy documents, regardless of working memory ability. Since higher education is often one's first exposure to heavily-mathematical documents, through lowering the technical barrier, we hope to enable a smoother technical transition for potential STEM majors.

2

Tool Design and Creation

This following chapter walks through the different components of the tool, from conception to code. The source files can be found on my Github. The sections below first describes the function of the tool before going into detail about design decisions and the various functions and data passed.

2.1 Platform

The initial proof-of-concept has been created as a Google Chrome Extension. Compared to the Mac app store or Adobe Reader add-ons, this platform has the greatest ease of creation and publishing, which allows for quick testing and turnaround time for iterations. Furthermore, Chrome is widely accessible across all laptop brands, thus allowing for easy distribution of the tool to participants.

2.2 Design

2.2.1 Overview

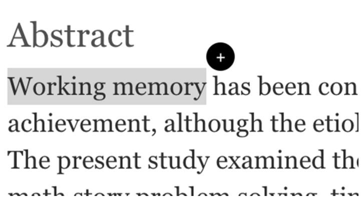

Figure 2.1: Initiating the creation flow

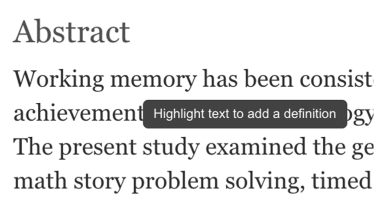

Figure 2.2: Input option

Figure 2.3: Overview of screen

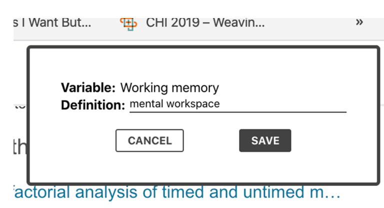

Figure 2.4: Close-up of card

The figures above illustrate the behavior of the tool. The user flow is as follows:

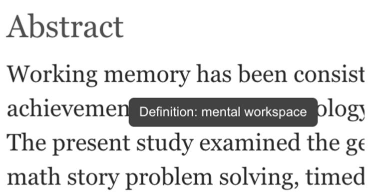

Figure 2.5: Resurfacing of definition

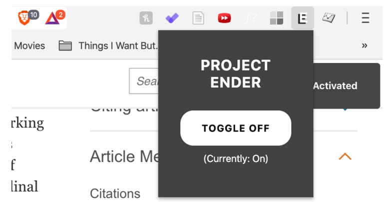

Figure 2.6: Toggle extension

1. Upon the user highlighting a letter, symbol, or phrase, a button appears to add the selected string as a saved variable (figure 2.1).

2. Clicking the button surfaces a card floating in the top-right of the window that shows the user the current state of the variable and definition that he or she wants to add (figure 2.4).

3. The user can then add a definition by either highlighting text (figure 2.3) or physically typing into the textbox before clicking 'submit' or pressing enter in order to save the variable.

4. Once the variable has been saved, hovering over any instance of the phrase on the page surfaces a tooltip that displays the definition assigned to the variable next to the cursor (figure 2.5).

5. This tool can be toggled on and off by opening the Extension display (figure 2.6).

It is important to emphasize that a simple and accessible interface is key to the performance of the tool. Otherwise, users would rather continue their original reading habits instead of investing additional energy to the learning and utilization of this tool. The following subsections describe the different considerations and principles used to guide design decisions.

2.2.2 Input

The process of adding a variable must be quick and seamless for the user to prefer the tool over mentally storing it in working memory. In order to minimize the friction encountered in defining a variable, the design of the variable addition flow was optimized along two metrics: noisiness and time.

In reference to noisiness, the flow for adding a new variable should not distract the reader to a point where it is more of a nuisance than an aid. To do this, the option to add a variable only surfaces at the proper times, and, in cases where it surfaced inappropriately, that it did not negatively impact the reader.

Therefore, as seen in figure 2.1, the add button only appears upon highlight of the text. When it does, it floats above and to the right of the highlighted text in order to avoid potentially blocking other characters. Any next mouseclick not on the button itself causes the add button to disappear, thereby minimizing the noise of its appearance when not needed. In addition, by requiring interaction with a switch to enable the extension (figure 2.6), this tool only surfaces on the pages that the user explicitly signals.

Secondly, this design aimed to minimize the time spent adding a variable and its definition. By surfacing the creation flow upon highlight of a word, selected word can automatically be set as the input variable. This removes the additional need for the user to add an input variable during the flow. Furthermore, through allowing the definition to be added not only through physically typing phrases but also highlighting text, the number of mouse and keyboard clicks were greatly reduced from the multiple keypresses needed to completely type out a word to a single additional highlight. The final creation flow only requires four clicks: one click to highlight a phrase to initiate the process, one on the '+' button to set it as the variable, one to highlight a phrase to input a definition for the variable, and a click on submit or keypress on enter to save the entry.

2.2.3 Displays

In line with the minimizing the negative externalities of the tool's presence, the UI elements of the display were modified to be as least invasive as possible. While the informational card (figure 2.4) originally hovered near the text set as the input variable, this was extremely intrusive and blocked a lot of potentially relevant text. By floating it on the top right of the screen, the main text area remained undisturbed, while also setting expectations for the position of the card, making it easier to find in subsequent creation flows.

However, by Fitts's law[7], the time it takes to click a target is proportional to the distance and size of the target; therefore, by moving the card so far away, this greatly increased the time it took for a user to type into and press submit on the card. Therefore, as seen in figure 2.2, I also allowed for additional inputs, such as highlighting words to add as definition and pressing enter to submit, to require as little movement as possible. In order to increase understanding of the flow, tooltips were added in the intermediary steps to explain the controls available.

When looking at resurfacing previously-entered definitions upon hover over a variable instance, I aimed to display the definition as close as possible to the mouse, as this is where the user's current focus was. By the Gestalt principle of proximity[8], since the definition is physically close to the initial variable instance, they are more likely to be interpreted as a single object, thereby contributing to the process of variable-definition mapping and understanding.

2.3 Code and Creation

The script files are broken into two portions, the extension interface (popup) and the content script listening in the background of the active tab. Data is passed from the content script to the extension interface, which then toggles the appropriate state. All visualizations are created in the content script.

2.3.1 Extension Interface

The extension interface, upon opening, queries the content script about the state of the switch, and modifies its display accordingly. It then listens for any interaction with the switch to signal the content script about whether to turn the tool on or off.

2.3.2 Content Script

The content script can be broken down into three functions: initializations, handling page interactions, and saving the variables.

In the initialization function, the script pulls the saved variables and switch state from storage and opens a listener for additional messages from the extension interface.

The function that handles page interactions is the meat of the script. This function, after every mouse click, evaluates the current state of the page. It funnels the input into two cases: starting the creation flow, and finishing the flow. If no variable has been defined, the '+' sign to add a variable (figure 2.1 is inserted to the DOM, along with a listener. Upon detection that it has been clicked, the remaining UI elements, such as the card and tooltip, are injected into the DOM. If the handler detects that the submit button or enter key has been pressed, it then calls the save function. Otherwise, if the mouse clicks away from the '+' sign, the creation flow is cancelled and the DOM elements removed.

The save function saves the variable to local storage and removes the superfluous DOM elements. It additionally scans through the body of the document and searches for instances of the variable. Since text elements cannot be detected by themselves without a DOM element, upon finding an instance of the variable, the function wraps the instance in a span, so that, upon hovering over the element, the tooltip is fired with the definition. The DOM wrapping function has been taken from an open source file by James Padolsey.

2.4 Limitations

Although Google Chrome allows for efficient update and creation of the tool, it has not been without its limitations. Specifically, Chrome does not allow for script injections into existing scripts, thereby preventing us from interaction from its native PDF viewer. Therefore, this extension only works on pure HTML webpages. While this only adds a slight complication to the proof of concept, this is a large limitation for actual implementation and use of the tool.

In order to prevent reference of variables when responding to questions, the storing of variables across webpages was disabled for the duration of the experiment.

Furthermore, the proof of concept as it stands only listens for UTF-8 characters; therefore, in the case that equations appear as images or contain symbols outside of this characterset, they will not be able to be recognized.

3

Methodology and Experimental Design

This chapter walks through the design and execution of the experiment.

3.1 Overview

This experiment, nicknamed Project Ender, aims to examine the efficacy of this tool in improving understanding of mathematical texts. A within-subjects test was conducted both in-person and remotely over 45 Harvard students concentrating in subjects that often encountered mathematical texts, such as STEM and Economics. During this experiment, participants read through two different mathematical texts, one with the opportunity to use the tool and one without, before answering comprehension questions related to the text they just read. All users participated in both conditions, with the order of the texts and tool randomized to mitigate novelty and practice effects.

3.2 Participants

46 Harvard students (22 male, 24 female) concentrating in mathematically rigorous concentrations (specifically, in STEM fields and/or Economics) were recruited for this study. Participants were recruited through email lists in the STEM community and through personal outreach to colleagues in STEM extracurriculars and courses. Aside from the provision of snacks, no other incentive was given. 14 (30%) participants indicated a concentration or secondary in Mathematics, 12 (26%) in Economics, 28 (61%) in Computer Science, and 12 (26%) in other categories (statistics, engineering, neuroscience). Given that most first-years have not decided on their concentration, there is an upperclassmen skew to the grade-level of participants, with only 4 (8%) freshmen, compared to 16 (35%) sophomores, 10 (22%) juniors, and 16 (35%) seniors.

Participants were also asked two control questions regarding how many pages and how many minutes a week they spent reading text of a similar mathematical rigor. For pages a week, the mean was 16.2, with a standard deviation of 27.5. For minutes a week, the mean was 89.2, with a standard deviation of 139. 17 students (37%) reported zero pages and zero minutes a week reading similar text. However, it is important to note that, as 35% of participants were seniors in their last semester, it could be the case that these students were not taking mathematically rigorous courses in the present, but have in the past during their time at Harvard. No information was collected about participants' overall relationship with mathematical texts throughout their entire course in college.

3.3 Pre-task

Before the session, participants were asked to download the Chrome extension and load it onto their computers. They were further instructed to close all other windows and silence their cellphones to eliminate distractions throughout the study. After being briefed about the survey they would encounter, participants then navigated to bit.ly/projectender and began the self-driven study. The full instructions can be found in the Appendix A.1.

3.4 Onboarding and Tutorial

Before starting the tasks, participants underwent a quick tutorial of the tool in order to build familiarity with using it. The tutorial consisted of six pages and three parts: selecting the variable, assigning it a definition, and revisiting the variable.

1. **Selecting the variable:** Instructions were given to highlight a word and press the 'plus' button in order to select it as a variable to save.

2. **Assigning a definition:** Participants were taught to do this in two ways: through physically typing a definition in the interface, or through highlighting another word or phrase to automatically populate the field. They were then taught to save by either pressing [Enter] or clicking the 'save' button.

3. **Revisiting the variable:** Instructions indicated that definitions surfaced upon hovering over any instance of the saved variable. Further clarification was given about the case-sensitivity of the variables.

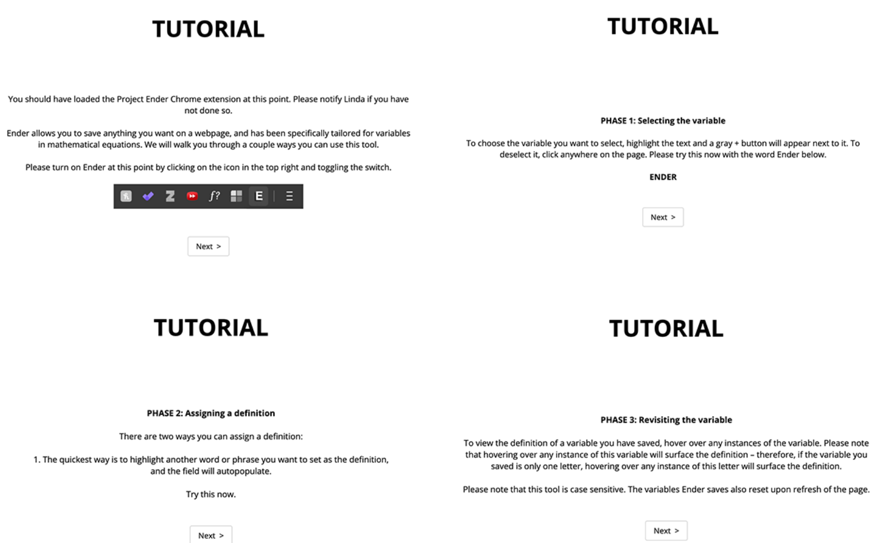

Figure 3.1: Selected views from the tutorial and walkthrough

During each phase, participants were given time to execute the instructions and familiarize themselves with the tool before proceeding.

3.5 Tasks

In both conditions, participants were asked to read a passage of mathematical text for eight minutes and were instructed to focus solely on the text during the duration of the countdown. After eight minutes, the page auto-advanced into a waiting screen that no longer contained any information pertinent to the passage. They were then instructed to navigate to the Qualtrics survey and answer some questions about the passage they had just read.

3.5.1 Passages

As the performance of the tool would be analyzed in relation to the passages chosen, careful consideration along many axes was given when searching for mathematical texts to use. Specifically, a balance had to be found between difficulty and length. The text could not be too basic that participants would have already known the content beforehand, but could not be too difficult that it relied on external knowledge that not all participants would have. As well, the text could not be too short, resulting in few variables and no need for the tool, but could not be too long in order to attract as many participants as possible and to minimize order effects of exhaustion and cognitive overload upon reaching the second passage.

While excerpts from textbooks were considered, due to permissions and difficulty finding short, isolated passages, the focus shifted to text from the public domain (specifically in Nobel Laureate works). Economics was chosen as the broad subject of focus as it was the most easily understandable for STEM concentrators of all backgrounds.

The passages ultimately chosen were excerpts from "The Market for Lemons" and "The Economics of Caste and of the Rat Race", both written by the same author, George Akerlof, in order to simulate as similar of an experience as possible between the two texts. Both texts have a length of three pages. The Lemons paper is slightly shorter, at 874 words, while the Caste paper is 50% longer, at 1302 words. Both excerpts have 12 variables each; however, the Lemons paper has 50% more equations (23 total) compared to the Caste paper (10 total). The full passages can be found in Appendix A.2.

3.5.2 Questions

Survey questions were administered after each reading, structured as follows:

- Five general demographic and experience questions examining:
  - Prior familiarity with the paper
  - How much of the paper was read within the time frame
  - How easy or difficult the participant found the paper
  - How much of the paper the participant understood
- Two summary questions about the passage asking:
  - The main argument
  - How the main argument was supported
- Five passage-specific questions about the passage looking at:
  - Understanding of the overarching content and pieces involved in the passage
  - Comprehension of the main argument and applicability to real-life
  - Ability to map understanding of the excerpt to another closely-related topic

– Recall of specific variables and equations encountered
– Recollection of the proof leading up to the equations

The full text of questions asked can be found in Appendix A.3.

Each question out of the five passage-specific questions were individually shown, as to not provide extra information that would give away part of the answer to another question. Participants were not allowed to backtrack and look at previous questions once they pressed continue.

These questions analyzed different aspects of comprehension: the summary questions, as well as the second passage-specific question, tested participants' abilities to create an overview of the passage. The first passage-specific question looked at participant's understanding of the overarching content. The third passage-specific question tested whether participants could map the ideas of the excerpt to another related area, while the fourth and fifth questions tested memory of very specific variables and equations encountered in the text.

In order to avoid participants specifically scanning through text to find answers, the questions were only provided after the eight minute reading period was over. However, since the questions for each passage followed a similar structure, there are novelty effects with the order of passages, as participants have a better idea of what to pay attention to the second time around. Therefore, passages were randomized so that econometric measures can be used to analyze and isolate the within-subject effects of the tool.

As time and length of responses could differ depending on understanding of the passages, there was no time limit imposed on the survey in order to provide participants room to respond as they see fit. There is a weak correlation (statistically significant at the 5% level) between time spent and scores on responses to questions 1, 2, 3 and 4 for the Lemons paper and questions 1, 2 and 5 for the Caste paper. However, there is also a statistically significant weak correlation (0.36) between time spent responding to the survey and participants' self-reported percent understanding of the Lemons passage, supporting the idea that a positive relationship exists between understanding of the passages and time spent answering the questions.

More importantly, as seen below, there is no statistically significant difference between time spent and the scores to responses with and without the Ender tool. Therefore, the upcoming analyses on difference-in-performance via usage of the tool do not factor in time spent answering the survey.

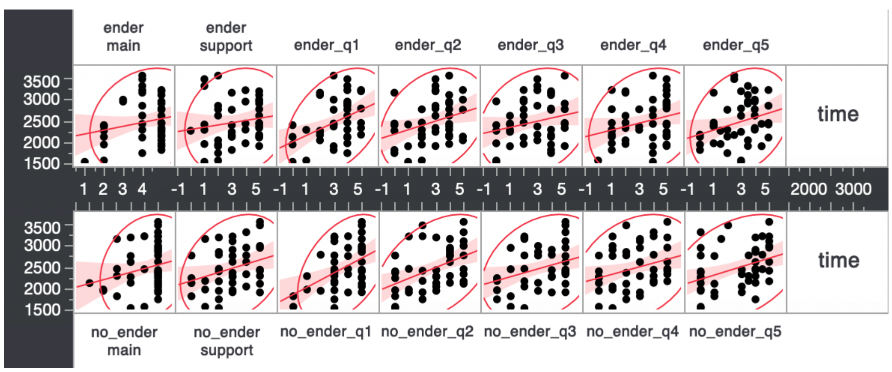

Figure 3.2: Correlation between time spent responding and scores for responses by tool usage

3.6 Survey Environment

While the study was originally planned to be fully conducted in-person in a specific, isolated room on campus, only thirty experiments were able to be finished there before the university was vacated due to external factors. Therefore, the last sixteen experiments were conducted remotely over Google Meets with specific instructions given prior to the study for participants to isolate themselves from external distractions for 45 minutes in order to replicate the test environment as much as possible. Aside from the instructions and environment, nothing else from the study (such as the experimental website and tool) was modified in any way when conducted remotely. No correlation (mean = 0.03, stddev = 0.12) was found between the study environment (in-person vs. remote) and the responses to the survey questions.

3.7 Data Processing

Using Amazon Mechanical Turk, responses to each of the eight questions for the two excerpts were rated along a scale of 0-5 from a given rubric. The full rubric can be found here, or in Appendix A.4.

Market for Lemons

What is the main argument of the paper?

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	n/a
1	[wrote something]
2	information symmetry,' 'information asymmetry' [or] market, buyers, sellers
3	symmetry is good [or] asymmetry is bad
4	(a)symmetry **affects** quality of goods (/cars/lemons) [or] quality of information
5	Asymmetry results in a market where no **trades occur**

Figure 3.3: Preview of grading rubric

For each question, specific responses were fed in through a csv file, and graders were asked to select the appropriate level of detail of the response, given the rubric.

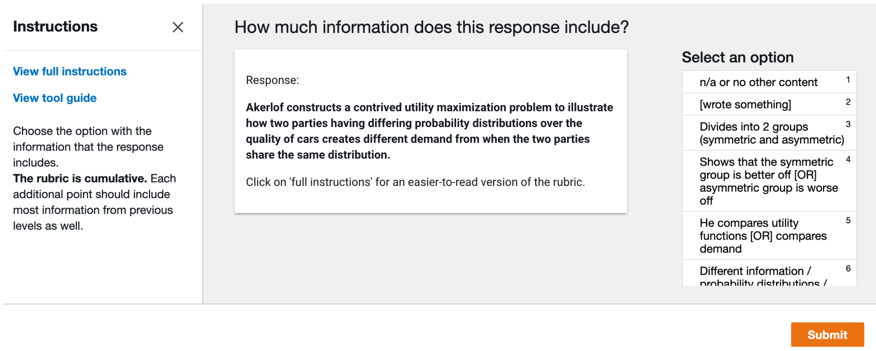

Figure 3.4: Example MTurk grading view

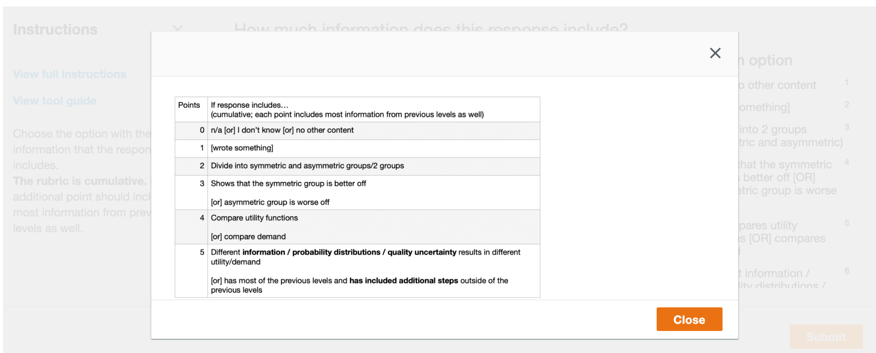

Figure 3.5: Clicking 'view full instructions' displays the full rubric

One limitation of the Mechanical Turk interface was the lack of text styling; while the rubric included words in bold to emphasize specific phrases to look out for, this did not translate onto the option selector. Due to this and the MTurkers' incentive to finish tasks as quickly as possible, it was clear for around 20% of responses that the MTurker did not follow the rubric; these obvious bad-actors were rejected. This number was especially large for rubrics that were very text-heavy. Furthermore, as no additional qualification types were specified in the MTurk batches, many of the graders were first-time MTurkers with very few prior approvals.

To increase robustness against bad actors, each response was graded three times and the

47% (18/38)	N/A	Different information / probability distributions / quality uncertainty results in different utility/demand [OR] [has included additional steps outside of the previous levels]
15% (6/41)	N/A	Shows that the symmetric group is better off [OR] asymmetric group is worse off
84% (32/38)	N/A	Different information / probability distributions / quality uncertainty results in different utility/demand [OR] [has included additional steps outside of the previous levels]

Figure 3.6: Example of rejected responses. The columns represent the MTurker's approval rate, the provided response, and their choice of grade from the rubric.

median taken to represent the final grade. The median was taken instead of the mean in order to ensure that outliers from undetected bad actors did not influence the final values.

The original table of responses (found here) was then translated to a table of numerical values for each response (found here). This table has 45 columns, as follows:

1. Number of Ender uses (numerical)
2. Order of the Lemons session (0 or 1)
3. Whether Ender was used in the Lemons session (boolean)
4. How familiar the participant was with the Lemons paper (1-5)
5. Whether the participant has read the Lemons paper before (boolean)
6. How much the participant had completed of the Lemons reading (1-7)
7. How difficult they found the Lemons reading (1-5)
8. What percentage of the Lemons reading they understood (0-100)

9. Score for Lemons main argument question (0-5)
10. Score for Lemons main support question (0-5)
11. Score for Lemons question 1 (0-5)
12. Score for Lemons question 2 (0-5)
13. Score for Lemons question 3 (0-5)
14. Score for Lemons question 4 overall (0-5)
15. Score for Lemons question 4 part 1 (0-3)
16. Score for Lemons question 4 part 2 (0 or 1)
17. Score for Lemons question 4 part 3 (0 or 1)
18. Score for Lemons question 5 overall (0-5)
19. Score for Lemons question 5 part 1 (0-2)
20. Score for Lemons question 5 part 2 (0-1)
21. Score for Lemons question 5 part 3 (0-1)
22. Score for Lemons question 5 part 4 (0-1)
23. Order of the Lemons session (0 or 1)
24. Whether Ender was used in the Caste session (boolean)
25. How familiar the participant was with the Caste paper (1-5)

26. Whether the participant has read the Caste paper before (boolean)
27. How much the participant had completed of the Caste reading (1-7)
28. How difficult they found the Caste reading (1-5)
29. What percentage of the Caste reading they understood (0-100)
30. Score for Caste main argument question (0-5)
31. Score for Caste main support question (0-5)
32. Score for Caste question 1 (0-5)
33. Score for Caste question 2 (0-5)
34. Score for Caste question 3 (0-5)
35. Score for Caste question 4 overall (0-5)
36. Score for Caste question 4 part 1 (0-3)
37. Score for Caste question 4 part 2 (0 or 1)
38. Score for Caste question 4 part 3 (0 or 1)
39. Score for Caste question 5 (0-5)
40. Gender (M, F, other)
41. Graduation year (2020 - 2023)
42. Concentration (text)

43. Secondary (if applicable)

44. How many minutes of similar text the participant reads a week (numerical)

45. How many pages of similar text the participant reads a week (numerical)

4

Results

THIS CHAPTER ANALYZES THE DATA collected from the experiment, starting at summary statistics and then examining in-detail the results for each of the eight questions, as well by background levels. The MTurk data was grouped by responses with-and-without Ender for the analysis. This new table can be found here.

4.1 Hypotheses

Before conducting the experiment, the following hypotheses were made:

- *H1.* The Ender tool will increase efficiency of reading through notation-heavy text.
- *H2.* The Ender tool will improve understanding of equations encountered in notation-heavy text.
- *H3.* Through improving understanding of equations, the Ender tool will increase overall comprehension of notation-heavy text.
- *H4.* The Ender tool will have a greater effect on those who have less experience reading through notation-heavy text.

4.2 Overview

Summary statistics for **responses with Ender** are below. 'percent_understand' refers to the self-reported percentages of understanding for each passage. 'main_argument' and 'main_support' are the scores for responses to the questions asking about the author's main arguments and how he supports them in detail, respectively. Questions 1-5 refer to the passage-specific questions asked, each tackling a slightly different facet of understanding.

	Percent Understand	Main Argument	Main Support
Median	50	5	3.5
Mean	51.21	4.13	3.35
StdDev	16.62	1.19	1.57

	Question 1	Question 2	Question 3	Question 4	Question 5
Median	3.5	3	3	4	3
Mean	3.09	2.78	2.74	3.33	2.89
StdDev	1.49	1.59	1.65	1.66	1.50

Summary statistics for **responses without Ender** are below.

	Percent Understand	Main Argument	Main Support
Median	50	5	3
Mean	48.61	4.09	2.98
StdDev	20.02	1.38	1.57

	Question 1	Question 2	Question 3	Question 4	Question 5
Median	3.5	3	3.5	3	3.25
Mean	3.59	2.80	3.15	2.59	2.60
StdDev	1.36	1.76	1.76	1.78	1.98

Looking these summary statistics, the values for the main overview responses using Ender are slightly higher, as well as Questions 4 and 5, are slightly higher than their non-Ender counterparts.

4.3 Controls

In the demographic questionnaire at the end of the survey, two control questions were asked about the number of minutes and pages spent reading text of similar mathematical rigor. While there was no significant correlation between minutes/pages read per week and performance on the questions using the Ender tool (m = 0.08, sd = 0.04 and m = 0.09, sd = 0.12, respectively), weak correlation was found between prior experience and performance on questions where the Ender tool was not permitted. Although most questions did not exhibit correlation for minutes/pages read and no-Ender results (m = 0.21, sd = 0.10 and m = 0.17, sd = 0.11, respectively), there was a significant (but weak) correlation of 0.3631 for question 4 between minutes spent reading and performance and 0.3617 for the question asking about Akerlof's support of his argument between pages read and performance. This correlation makes sense given the context that question 4 asked about specific equations and variables from the passage and the supporting evidence question also required very specific knowledge from the passage. Those who are accustomed to reading similar text may have more experience with internalizing minute details of passages, especially in this case where the Ender tool is not present to act as a crutch for this.

Below are summary statistics for the general background questions.

	Ender Session (0,1)	Ender Paper (0 = C, 1 = L)	Number of Ender Uses
Median	1	1	7.5
Mean	0.65	0.65	7.41
StdDev	0.48	0.48	3.80

Table 4.1: Summary statistics for Ender use

While the website chose the order of text and order of Ender tool use at random, the groups are not completely balanced; there is a slight skew in Ender use toward the first session and the Lemons paper. The number of Ender uses ranged greatly, from zero (even with the tool enabled) to fifteen saved variables.

	Familiarity with Ender Paper (0-5)	Ender Paper Amount Completed (0-7))	Ender Paper Difficulty (0-5))
Median	2	5	2
Mean	1.76	5.13	2.48
StdDev	0.90	1.59	0.89

Table 4.2: Summary statistics for overview questions for passages with Ender

	Familiarity with No-Ender Paper (0-5)	No-Ender Paper Amount Completed (0-7))	No-Ender Paper Difficulty (0-5))
Median	1	5	2
Mean	1.32	4.93	2.35
StdDev	0.70	1.73	0.87

Table 4.3: Summary statistics for overview questions for passages without Ender

The responses for the excerpts without Ender use are, on average, slightly lower than those with Ender use. This could be due to the slight skew toward Ender use with the Lemons paper, which students from Applied Mathematics and Economics signalled more familiarity with from previous classes. Additionally, one participant indicated that they had read the Lemons paper prior to the study; however, no significant correlation was found between this and any of the participants' survey responses, so this factor was not taken into account in further analyses.

4.4 Analysis of Specific Responses

The following section dives deeper into a within-subjects analysis of each content question. A multivariate analysis of variance (MANOVA) test was run on each of the questions to analyze the within-subject difference of using the Ender tool, controlling for the within-subject passage and tool order and between-subject difference in prior experience with reading (in minutes and number of pages per week). Additionally, responses were analyzed in two cases: with and without weight by number of uses of the Ender tool. As certain participants were less familiar with the tool or did not find a need for it, there were cases where the tool was used infrequently (including one case where it was not used at all). This additional analysis aims to isolate the effect of actual usage of the Ender tool.

4.4.1 Percent of Perceived Understanding

There is an slight average within-subject difference in perceived understanding of 2.56 percent (at the 5% significance level, p-value = 0.042) when weighing responses by number of uses of the Ender tool. Without weighing, there is an average difference of 2.6 percent (at the 10% significance level, p-value = 0.095).

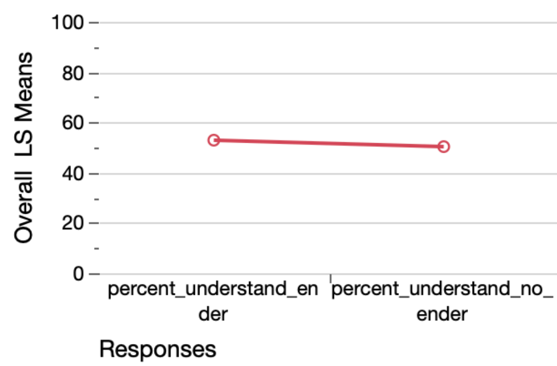

Figure 4.1: Average reported percent understanding, weighed by num tool uses

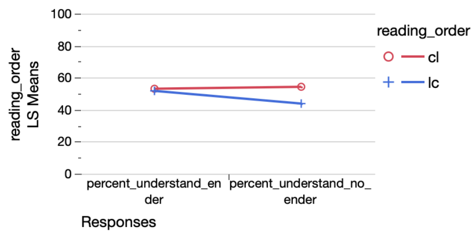

Figure 4.2: Avg understanding based on reading order

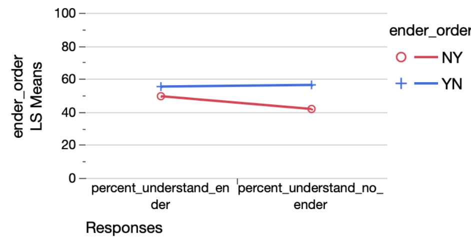

Figure 4.3: Avg understanding based on Ender order

4.4.2 Akerlof's Main Argument

No significant difference was found between grades for the main argument of the paper. The average difference was -0.018 (p-value = 0.980) when weighing by number of Ender uses, and 0.043 (p-value = 0.725) without weight.

4.4.3 Akerlof's Support of His Main Argument

No significant difference was found between grades for the question asking about the support of Akerlof's main argument. The average difference was 0.224 (p-value = 0.246) when weighing by number of Ender uses, and 0.37 (p-value = 0.185) without weight.

4.4.4 Question 1

Question 1 examined understanding of the overarching content and pieces involved in the passage. No significant difference was found between grades for this question – the average difference was -0.334 (p-value = 0.323) when weighing by number of Ender uses, and -0.5 (p-value = 0.197) without weight.

4.4.5 Question 2

Question 2 looked at comprehension of Akerlof's main argument through measuring applicability to real-life scenarios. No significant difference was found between grades for this question – the average difference was 0.117 (p-value = 0.359) when weighing by number of Ender uses, and -0.021 (p-value = 0.480) without weight.

4.4.6 Question 3

Question 3 measured participants' ability to map understanding of the excerpt to another closely-related topic. No significant difference was found between grades for this question – the average difference was -0.342 (p-value = 0.910) when weighing by number of Ender uses, and -0.413 (p-value = 0.813) without weight.

4.4.7 Question 4

Question 4 asked about specific details regarding equations and variables encountered in the passage. Weighing by number of times the Ender tool was used, there is an average difference of 0.650 (around 24%) in scores with-and-without access to the Ender tool, significant at the 5% level (p-value = 0.009). Without weighing by Ender uses, there is a difference of 0.739, with a p-value of 0.005.

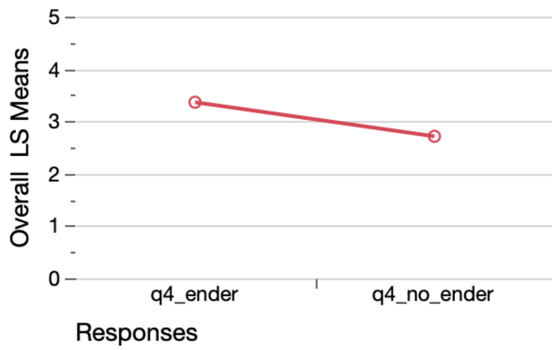

Figure 4.4: Average scores on question 4, weighed by num tool uses

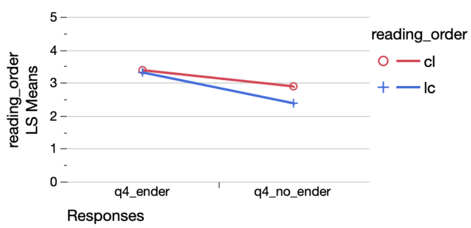

Figure 4.5: Avg q4 scores based on reading order

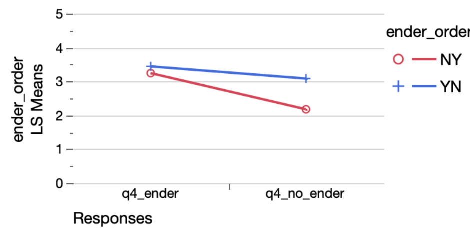

Figure 4.6: Avg q4 scores based on Ender order

4.4.8 Question 5

Question 5 also asked about equations, but in a larger context, focusing on recollection of the proof leading up to the equations. No significant difference was found between grades for this question – the average difference was 0.208 (p-value = 0.2267) when weighing by number of Ender uses, and 0.293 (p-value = 0.172) without weight.

4.5 Results by Background Level

While background level was controlled for in the analyses above, this section looks at whether differences exist in effect of Ender tool usage by amount of previous experience reading similar text. We examine the two questions where a significant difference was found by tool usage: percent understanding and question 4. Responses were split into two groups based on reported minutes and pages of similar text read per week. As there was no easy divide through the reports of smaller numbers, participants were split into two groups based on whether they reported reading any pages/minutes at all each week. Therefore, one group consisted of 18 participants who reported 0 pages and 0 minutes of reading notation-heavy text a week, while the other group consisted of the 28 other participants.

4.5.1 Background Level x Percent Understanding

There is a large and significant effect of usage of the Ender tool and self-reported understanding of the passage for participants with little background experience. Weighing by number of Ender uses, this is a 11.934 difference in percentage points (p-value = 0.018), and a 10.941 difference (p-value = 0.022) without weight.

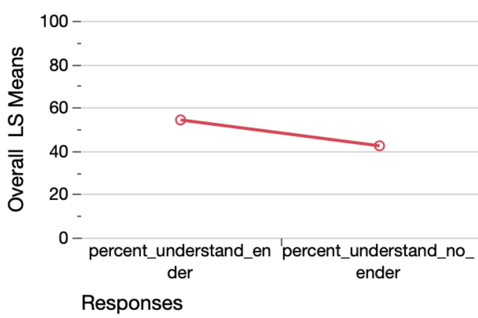

Figure 4.7: Pct understanding for those with no experience

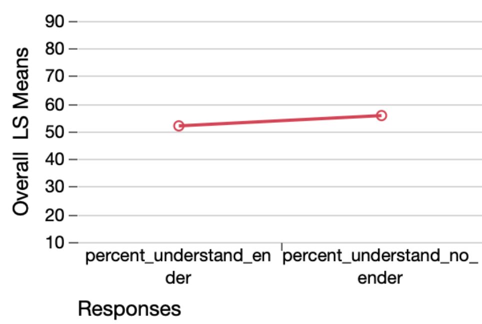

Figure 4.8: Pct understanding for those with some experience

Compared to the group with no experience, there is a negligible and insignificant difference of -3.73 percentage points (p-value = 0.403) for those with background experience, when weighed by Ender use, and -2.285 (p-value = 0.549) without weight.

4.5.2 Background Level x Question 4

Similarly, there is also a large and significant difference for the scores of question 4 for the no-experience group when the tool is used, with a difference of 1.357 (a 38% change, p-value = 0.080) when weighed by number of tool uses, and 1.471 (p-value = 0.038) when not weighed.

There is no significant result for the scores of the group with prior reading experience; there is a difference of 0.177 (p-value = 0.22) when weighed by amount of Ender use, and 0.310 (p-value = 0.145) without this weight.

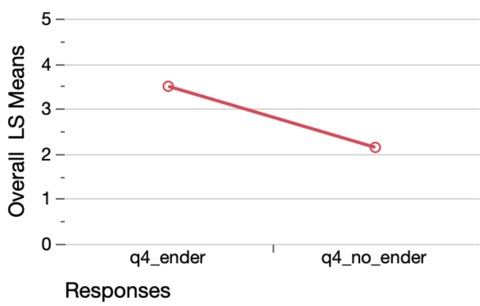

Figure 4.9: Q4 scores for those with no experience

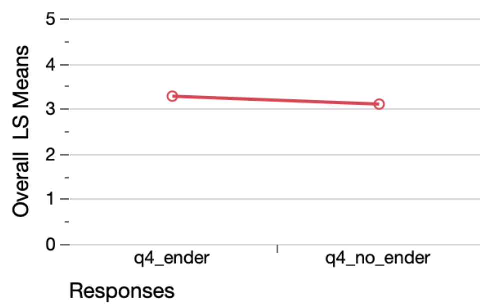

Figure 4.10: Q4 scores for those with some experience

To analyze the specific difference between low and high-experience, the 'some-experience' group was additionally split into two subgroups, one consisting of 14 participants who reported reading similar text for 60 minutes or less a week, and the other group consisting of 14 participants who reported reading for more than 60 minutes (ranging from 100 to 600 minutes) per week.

While the count is too small to draw significance (p-values of 0.983 and 0.280), it is interesting to highlight that the data shows diminishing returns for participants with greater background experience. As seen above, participants with no experience had an steep average increase of 1.357 on their Q4 scores when using Ender. Those with low-experience (1 hour or less a week) had a difference of 0.381 on their scores with Ender (figure 4.11), while those with high-experience had an opposite result, where Ender decreased their scores on average by -0.077 (figure 4.12).

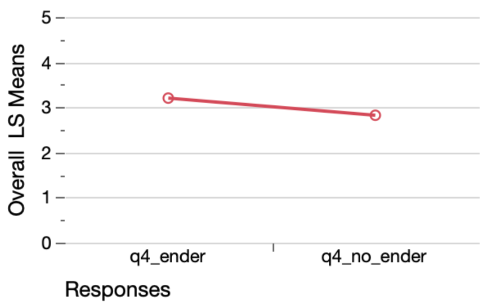

Figure 4.11: Q4 scores for those with low experience

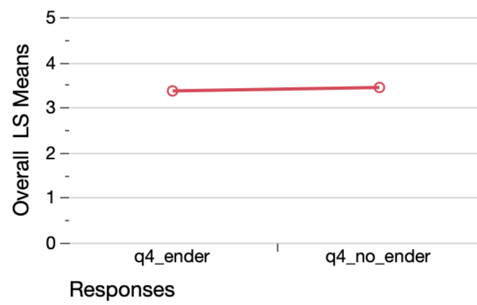

Figure 4.12: Q4 scores for those with high experience

4.6 Hypotheses Revisited

The first hypothesis posited that the Ender tool will increase the efficiency of reading through notation-heavy text. As the experiment set a standardized 8-minutes for reading through each excerpt, efficiency could not be measured through time spent in this case. However, using self-reported percent understanding of the excerpt as a proxy for reading speed, the data has shown a small but significant effect of usage of the Ender tool with understanding, supporting this hypothesis.

Hypothesis two posited that usage of Ender will improve understanding of equations encountered in notation-heavy text. The findings from question 4 suggest that the tool increased understanding of equations by an average of 24%, supporting this hypothesis.

Regarding hypothesis three, no significant difference was found in questions tackling conceptual knowledge of the passages. This is left inconclusive.

Hypothesis four postulated that there will a greater effect of tool use for those with less experience reading through notation-heavy text. While no significant results were found for high-experience participants, zero-experience participants had a significantly larger benefit (when from use of Ender compared to the average) in both questions that showed original significance. This hypothesis is supported specifically in the two cases of percent_understanding and question_four, but inconclusive for effects on overall comprehension.

5

Conclusion

5.1 Discussion and Implications

In an increasingly online world, especially in this current time when academics have moved virtually, the digitalization of text provides new opportunities to enhance learning and comprehension previously unavailable to paper copies. This experiment looks at how digital tools can be leverage to aid in the reading process. The results show a significant effect of usage of the Ender tool with understanding of passages, especially in regards to equation-specific details. Participants received, on average, a 20% increase in scores for questions asking variable and equation-specific details in passages where the Ender tool was used. This shows promise for methods that act as a crutch for working memory to increase understanding of mathematical equations in text around the 1,000 word length and involving 10-20 variables and equations. Reading efficiency, as measured through self-reported understanding of the passages, was increased by 5 percentage points on average in cases where the tool was used. Although this is a small difference, this result displays potential for aids in working memory to have an effect on reading efficiency and reader confidence.

While understanding in the self-reported and detail-specific contexts were improved by Ender, no significant results were found in the overall understanding of the passage nor in responses to conceptual questions. In addition, no relationship was found between usage of Ender and understanding the larger context of the proof leading up to the equations presented. It could be the case that detailed understanding does not easily translate to conceptual understanding or that, as related works suggest, working memory is more relevant in high-complexity situations and less in big-idea generalizations. The sample size is too small and has too many confounding factors to come to a conclusion in this case.

Perhaps the most salient finding is that the Ender tool had a much larger effect on participants with no background experience reading notation-heavy text of a similar caliber. For the questions focusing on equation-specific details, participants with no background experience saw a 1.357 point (50%) increase in scores in cases where Ender was used, which is an over-100% increase when compared to the general average increase to scores in question

4 with Ender. When self-reporting percentage understanding of the passage, participants with no background experience reported, on average, 12% greater understanding (a 20% increase) when Ender was used. This is a 9.374 difference and 300% increase, when compared to the population average with Ender. This highlights a unique opportunity to utilize tools such as Ender to level the playing field and increase the accessibility of mathematical texts for students with little background experience.

To conclude, Ender was created as a crutch for working memory, in hopes of improving understanding of mathematical text. Although no relationship was found in overall and bigger-picture comprehension, self-reported understanding of the text and detailed understanding of equations were improved through the usage of the tool, especially for participants with little background in similar text. As a computer science concentrator with no previous experience in this field until entering college, this study hits home and provides a glimpse of promise into lowering the steep learning curve when approaching notation-heavy text.

5.2 Further Steps

This section discussion future steps to expand upon results found in this study and addresses potential remedies to confounding factors found in this experiment.

5.2.1 Automation and NLP

While the results showed an increase in self-reported understanding, there was no significant effect (and an observed decrease) in general understanding of the passage when the

Ender tool was used, especially when weighed by number of uses. For example, the average difference in scores for the main argument question was -0.018 when weighed by Ender use, and 0.043 without weight. In this case, a trade-off could exist between usage of the tool and time spent, where the tool is able to promote specific understanding but the time spent adding definitions results in more difficultly extrapolating from the text.

This highlights a potential for natural language processing to decrease time spent and thereby improve efficiency and understanding. Through training on text with labelled variables, it is possible to extract features of variables and definitions that will enable easier, or even automatic, labelling of relevant equations.

5.2.2 Beyond University

While this study focused on a targeted group of college students, there is potential for this tool to be applied in multiple other use-cases where notation-heavy text is often encountered and detailed understanding is desired, such as academics who often read peer-reviewed papers or English literature where many characters or terms are introduced.

5.2.3 Choice of Passages

These conclusions made can only be safely extrapolated to text of a similar subject, length, and complexity as Akerlof's excerpts (Economics, 4-pages, 10-20 variables and equations, respectively). Future studies of a similar format can be conducted to explore the bounds of length and complexity in the effect of Ender.

5.2.4 Robust Grading

The responses were listed on Amazon Mechanical Turk for minimal payment ($0.05), which, coupled with a complex rubric, could have led to inaccurate grading and bad actors. Although this was controlled for as much as possible, a larger amount of listings with basic MTurk qualifications set would limit the potential for inaccuracy of grading. Additionally, a simplified rubric with exact phrases to match would make grading more quantitative, instead of providing room for error due to generalizations.

5.2.5 Participants

Only surface level demographic information was collected from participants. More fine-grained analysis could have been completed if more information were collected, such as overall reading habits throughout participants' time at Harvard, as well as tests to measure their working memory capabilities.

As well, insignificance could have resulted from the small sample size of 46 students, especially when broken down into further subgroups. A larger sample size could result in new significance for some findings.

5.2.6 Recall vs. Memorization

The survey in the experiment was administered right after the text was read, which tests participants' recall ability more than their long-term memorization. Studying the use of a tool in aiding working memory for re-reads of text or longer time intervals between a need to recall could show different results in its efficacy. In cases where students only revisit text much later in the semester, such as for exams, further research must be done to analyze how time effects Ender's performance.

A

Appendix

A.1 Full Experiment Instructions

The full instructions read to participants before conducting the experiment is as follows:

> Thank you for volunteering to participate in this experiment. Please take this moment to take out any digital devices on you, including but not limited to

phones, watches, and pagers, silence them, and put them away in your bag where they cannot be seen. [As you can see, this study is being conducted remotely. In order to replicate an in-person experience as much as possible, please try to solely focus on this for the next thirty minutes with as minimal distractions as possible. Please close or minimize all other tabs, including Facebook and Messenger, to eliminate distractions throughout the study.] This is a group study, but the experiments will be conducted individually. In order to eliminate potential distractions, please treat this study as if the other person and I are not here. I will proctor the study and debug any potential issues, should they arise, but otherwise cannot communicate nor answer any questions until after the study is complete.

Please take out your laptops at this time if you have not already. Please close or minimize all other tabs, including Facebook and Messenger, to eliminate distractions throughout the study. Please mute your laptop volume.

Navigate to bit.ly/projectender. During the course of the experiment, it is important that you do not refresh at any point to ensure smooth progression of the study.

You will encounter a survey during the study that will ask you some questions about readings. The structure of the survey is one general page, followed by five more specific questions on the passage, each on their own page. While you are encouraged to answer the questions thoroughly, if you are unclear how to respond, feel free to put N/A or stop answering when you do not have anything else to write. (This study should finish in 30 minutes but can take longer if you dally in filling out the survey.)

This study is self-driven. Please read through each page carefully, and click next when you are ready to proceed.

The sentences in brackets were added on for the studies that were conducted remotely. The paragraph about the survey was included to set expectations about the Qualtrics form after an initial test showed that participants struggled with the longer length of the survey.

A.2 Passage Excerpts

The excerpts from Akerlof's papers are found below. The "Market for Lemons" passage begins from section B, and the "Economics of Caste and of the Rat Race" passage begins at "Formal Model of Caste Equilibrium".

II. THE MODEL WITH AUTOMOBILES AS AN EXAMPLE

A. *The Automobiles Market*

The example of used cars captures the essence of the problem. From time to time one hears either mention of or surprise at the large price difference between new cars and those which have just left the showroom. The usual lunch table justification for this phenomenon is the pure joy of owning a 'new' car. We offer a different explanation. Suppose (for the sake of clarity rather than reality) that there are just four kinds of cars. There are new cars and used cars. There are good cars and bad cars (which in America are known as 'lemons'). A new car may be a good car or a lemon, and of course the same is true of used cars.

The individuals in this market buy a new automobile without knowing whether the car they buy will be good or a lemon. But they do know that with probability q it is a good car and with probability $(1-q)$ it is a lemon; by assumption, q is the proportion of good cars produced and $(1-q)$ is the proportion of lemons.

After owning a specific car, however, for a length of time, the car owner can form a good idea of the quality of this machine; i.e., the owner assigns a new probability to the event that his car is a lemon. This estimate is more accurate than the original estimate. An asymmetry in available information has developed: for the sellers now have more knowledge about the quality of a car than the buyers. But good cars and bad cars must still sell at the same price—since it is impossible for a buyer to tell the difference between a good car and a bad car. It is apparent that a used car cannot have the same valuation as a new car—if it did have the same valuation, it would clearly be advantageous to trade a lemon at the price of new car, and buy another new car, at a higher probability q of being good and a lower probability of being bad. Thus the owner of a good machine must be locked in. Not only is it true that he cannot receive the true value of his car, but he cannot even obtain the expected value of a new car.

Gresham's law has made a modified reappearance. For most cars traded will be the 'lemons,' and good cars may not be traded at all. The 'bad' cars tend to drive out the good (in much the same way that bad money drives out the good). But the analogy with Gresham's law is not quite complete: bad cars drive out the good because they sell at the same price as good cars; similarly, bad money drives out good because the exchange rate is even. But the bad cars sell at the same price as good cars since it is impossible for a buyer to tell the difference between a good and a bad car; only the seller knows. In Gresham's law, however, presumably both buyer and seller can tell the difference between good and bad money. So the analogy is instructive, but not complete.

B. *Asymmetrical Information*

It has been seen that the good cars may be driven out of the market by the lemons. But in a more continuous case with different grades of goods, even worse pathologies can exist. For it is quite possible to have the bad driving out the not-so-bad driving out the medium driving out the not-so-good driving out the good in such a sequence of events that no market exists at all.

One can assume that the demand for used automobiles depends most strongly upon two variables—the price of the automobile p and the average quality of used cars traded, μ, or $Q^d = D(p, \mu)$. Both the supply of used cars and also the average quality μ will depend upon the price, or $\mu = \mu(p)$ and $S = S(p)$. And in equilibrium the supply must equal the demand for the given average quality, or $S(p) = D(p, \mu(p))$. As the price falls, normally the quality will also fall. And it is quite possible that no goods will be traded at any price level.

Such an example can be derived from utility theory. Assume that there are just two groups of traders: groups one and two. Give group one a utility function

$$U_1 = M + \sum_{i=1}^{n} x_i$$

where M is the consumption of goods other than automobiles, x_i is the quality of the ith automobile, and n is the number of automobiles.

Similarly, let

$$U_2 = M + \sum_{i=1}^{n} 3/2x_i$$

where M, x_i, and n are defined as before.

Three comments should be made about these utility functions: (1) without linear utility (say with logarithmic utility) one gets needlessly mired in algebraic complication. (2) The use of linear utility allows a focus on the effects of asymmetry of information; with a concave utility function we would have to deal jointly with the usual risk-variance effects of uncertainty and the special effects we wish to discuss here. (3) U_1 and U_2 have the odd characteristic that the addition of a second car, or indeed a kth car, adds the same amount of utility as the first. Again realism is sacrificed to avoid a diversion from the proper focus.

To continue, it is assumed (1) that both type one traders and type two traders are von Neumann-Morgenstern maximizers of expected utility; (2) that group one has N cars with uniformly distributed quality x, $0 \leqq x \leqq 2$, and group two has no cars; (3) that the price of 'other goods' M is unity.

Denote the income (including that derived from the sale of automobiles) of all type one traders as Y_1 and the income of all type two traders as Y_2. The demand for used cars will be the sum of the demands by both groups. When one ignores indivisibilities, the demand for automobiles by type one traders will be

$$D_1 = Y_1/p \qquad \mu/p > 1$$
$$D_1 = 0 \qquad \mu/p < 1.$$

And the supply of cars offered by type one traders is

$$S_2 = pN/2 \qquad p \leqq 2 \tag{1}$$

with average quality

$$\mu = p/2. \tag{2}$$

(To derive (1) and (2), the uniform distribution of automobile quality is used.)

Similarly the demand of type two traders is

$$\begin{aligned} D_2 &= Y_2/p & 3\mu/2 &> p \\ D_2 &= 0 & 3\mu/2 &< p \end{aligned}$$

and

$$S_2 = 0.$$

Thus total demand $D(p, \mu)$ is

$$\begin{aligned} D(p, \mu) &= (Y_2 + Y_1)/p & &\textit{if } p < \mu \\ D(p, \mu) &= Y_2/p & &\textit{if } \mu < p < 3\mu/2 \\ D(p, \mu) &= 0 & &\textit{if } p > 3\mu/2. \end{aligned}$$

However, with price p, average quality is $p/2$ and therefore at no price will any trade take place at all: in spite of the fact that *at any given price* between 0 and 3 there are traders of type one who are willing to sell their automobiles at a price which traders of type two are willing to pay.

C. *Symmetric Information*

The foregoing is contrasted with the case of symmetric information. Suppose that the quality of all cars is uniformly distributed, $0 \leqq x \leqq 2$. Then the demand curves and supply curves can be written as follows:

Supply

$$\begin{aligned} S(p) &= N & p &> 1 \\ S(p) &= 0 & p &< 1. \end{aligned}$$

And the demand curves are

$$\begin{aligned} D(p) &= (Y_2 + Y_1)/p & p &< 1 \\ D(p) &= (Y_2/p) & 1 < p &< 3/2 \\ D(p) &= 0 & p &> 3/2. \end{aligned}$$

In equilibrium

$$p = 1 \qquad \text{if } Y_2 < N \tag{3}$$

$$p = Y_2/N \qquad \text{if } 2Y_2/3 < N < Y_2 \tag{4}$$

$$p = 3/2 \qquad \text{if } N < 2Y_2/3. \tag{5}$$

If $N < Y_2$ there is a gain in utility over the case of asymmetrical information of $N/2$. (If $N > Y_2$, in which case the income of type two traders is insufficient to buy all N automobiles, there is a gain in utility of $Y_2/2$ units.)

Finally, it should be mentioned that in this example, if traders of groups one and two have the same probabilistic estimates about the quality of individual automobiles—though these estimates may vary from automobile to automobile—(3), (4), and (5) will still describe equilibrium with one slight change: p will then represent the expected price of one quality unit.

III. EXAMPLES AND APPLICATIONS

A. *Insurance*

It is a well-known fact that people over 65 have great difficulty in buying medical insurance. The natural question arises: why doesn't the price rise to match the risk?

Our answer is that as the price level rises the people who insure themselves will be those who are increasingly certain that they will need the insurance; for error in medical check-ups, doctors' sympathy with older patients, and so on make it much easier for the applicant to assess the risks involved than the insurance company. The result is that the average medical condition of insurance applicants deteriorates as the price level rises—with the result that no insurance sales may take place at any price.[1] This is strictly analogous to our automobiles case, where the average quality of used cars supplied fell with a corresponding fall in the price level. This agrees with the explanation in insurance textbooks:

> Generally speaking policies are not available at ages materially greater than sixty-five. . . . The term premiums are too high for any but the most pessimistic (which is to say the least healthy) insureds to find attractive. Thus there is a severe problem of adverse selection at these ages.[2]

The statistics do not contradict this conclusion. While demands for health insurance rise with age, a 1956 national sample survey of 2,809 families with 8,898 persons shows that hospital insurance coverage drops from 63 per cent of those aged 45 to 54,

[1] Arrow's fine article, 'Uncertainty and Medical Care' (*American Economic Review*, Vol. 53, 1963), does not make this point explicitly. He emphasizes 'moral hazard' rather than 'adverse selection.' In its strict sense, the presence of 'moral hazard' is equally disadvantageous for both governmental and private programs; in its broader sense, which includes 'adverse selection,' 'moral hazard' gives a decided advantage to government insurance programs.

[2] O. D. Dickerson, *Health Insurance* (Homewood, Ill.: Irwin, 1959), p. 333.

Why should these three conditions describing marriage customs in India be of interest to the economist? First, note that those who fail to follow, or even to enforce the caste customs do not gain the profits of the successful arbitrageur but instead suffer the stigma of the outcaste. If the punishment of becoming an outcaste is predicted to be sufficiently severe, the system of caste is held in equilibrium irrespective of individual tastes, by economic incentives; the predictions of the caste system become a self-fulfilling prophecy.[23]

Second, the recent extensions of the model of supply and demand to discrimination, household organization, crime and marriage show that the boundaries between sociology and economics are by no means clear; if economic models can explain sociological phenomena, so also the process can work in reverse with sociological models describing economic phenomena. With appropriate adjustment, the model of marriage in India explains both economies pathologically different from the A-D utopia, and also special pathologies in economies in which perfect competition, or slight deviations therefrom, are the norm.

Finally, the formal model of caste equilibrium works spontaneously without direction of any individual or organization. But in this model it is also natural to have the exact same economic structure with some arbiter of the caste code. Indeed the model is therefore useful in indicating how individuals and organizations can yield great powers—quite possibly, as in some of the later examples, with considerable abuse.

Formal Model of Caste Equilibrium

This subsection presents a formal model of caste equilibrium. Caste equilibrium is defined as a state of the economy in which caste customs are obeyed, yet no single individual, by behaving differently, can make himself better off. The first concern is, of course, to describe this equilibrium. However, since there are also coalitions of individuals who by acting together can make themselves better off than in equilibrium, it is also of interest to know the relative ease or difficulty of forming such a coalition. For this purpose we also look at the size and nature of the smallest equilibrium-breaking coalition.

Four sets of assumptions describe the economy; those describing technology, market structures, tastes, and the social system. The assumptions describing the social system are laid out in parallel with the earlier description of marriage in India. In general this model is extremely simple, subject to one complication. By its very nature the caste system involves trade and the division of labor. If outcastes could set up their own economy independent of caste members, the caste system would fall apart. Therefore, three assumptions are inserted that lead individuals to trade with one another; laborers can produce only one product; firms produce only one product; and tastes are such that persons will wish to purchase more than one good.

[23] Note that this is the 'terrorist' model of economic activity. One good example is the terrorist regime of Henry V of England, described by G. Mattingly, *Catherine of Aragon* (New York: Random House Paperback, 1960). Note also that this model describes the college 'honors' systems.

*Technology T*1. There are three types of jobs: skilled jobs, unskilled jobs, and scavenging jobs. (Subscripts *sk, u*, and *sc* refer to *sk*illed, *un*skilled and *sc*avenging, respectively.)

*T*2. There are n different products, labeled $i = 1, \ldots, n$.

*T*3. The production of each product depends upon the quantity of labor employed and the jobs performed by the labor. Let θ_{sk}, θ_u, and θ_{sc} denote the output of one unit of labor in producing any product in a skilled job, unskilled job, or scavenging job, respectively. The production function of good i is then

$$q_i = \sum_j \theta_j n_{ij},$$

where

$j = sk,\ u,\ sc,\ i = 1,\ \ldots,\ n$

q_i = output of product i, and

n_{ij} = quantity of labor employed in job type j in production of good i.

Of course,

$$\theta_{sc} < \theta_u < \theta_{sk}. \tag{1}$$

*T*4. Because of economies due to specialization workers can work on the production of only one product.

Market Structures. All firms are competitive profit maximizers. These firms can produce only one product. They hire labor and sell output on the market. A firm is willing to bid for labor the expected marginal value product of that labor.

Tastes. All persons have the same utility function U, which is independent of the caste code.

$$U = \sum_{i=1}^{n} \min(x_i, \alpha), \tag{2}$$

where x_i is consumption of good i and α is a parameter of the utility function.

*Social Structure S*1. By birth there are just two castes divided into a dominant caste D and a nondominant caste N. Labor of both castes D and N can be outcasted. Outcastes, if any, form a third group.

*S*2. The caste code dictates that D labor may work in only skilled jobs; N labor may work in only unskilled jobs; and outcaste labor may only hold scavenging jobs. The caste code also says that all persons who purchase from firms not using labor according to the caste code will themselves be outcasted.

*S*3. Persons predict that breakers of the caste code will be outcasted and receive the wages bid for outcaste labor.

Caste Equilibrium. Let the economy be described as above. Let w_k, $k = D, N$ denote the wage of caste k. Let p_i denote the price of good i produced by firms that use labor according to the caste code. Let good 1 be the numeraire good, with price equal to 1. Assume parameter values

$$\alpha < (\theta_u - \theta_{sc})/(1 - \theta_{sc}/\theta_{sk}) \tag{3}$$

and

$$n > \theta_{sk}/\alpha. \tag{4}$$

The following describe an equilibrium with fulfilled expectations:

1. $w_D = \theta_{sk}$, $w_N = \theta_u$.
2. The price of all goods produced by firms using labor according to caste code is 1.
3. There are no outcastes. N labor works at unskilled jobs. D labor works at skilled jobs.
4. Utility of D labor is θ_{sk}; utility of N labor is θ_u.
5. The highest wage bid for outcaste labor is θ_{sc}.

A coalition of k^* firms, producing k^* different products and using outcaste labor in skilled jobs, can break this equilibrium if

$$k^* > (\theta_u - \theta_{sc})/\alpha(1 - \theta_{sc}/\theta_{sk}).$$

Proof. It is obvious that the described equilibrium is feasible. We need show only that no new firm can make zero or positive profits and bid a higher wage either for N labor or for outcaste labor.

N Labor. Suppose that a new firm bids a higher wage for N labor than θ_u. It must use some of this labor in skilled jobs. In this case its profits per laborer will not exceed

$$p\theta_{sk} - \theta_u,$$

where p is the price received for its product. If profits are nonnegative,

$$p \geqq \theta_u/\theta_{sk}.$$

But at a price as great as θ_u/θ_{sk} this firm will have no customers. Consider a prospective customer. This customer will be outcasted because N labor is used in skilled jobs. Therefore, his expected wage is θ_{sc}. He will maximize expected utility by purchasing α units at a price p and $(\theta_{sc} - \alpha p)$ units of other goods from other firms that use labor according to the caste code.

His total utility will therefore be

$$\theta_{sc} - \alpha p + \alpha \leqq \theta_{sc} - \alpha\theta_u/\theta_{sk} + \alpha. \tag{5}$$

But by (1) and (3) the right-hand side of (5) is less than θ_u.

Since the customer of this firm receives utility at least as large as θ_u if he does not purchase from the caste-breaking firm, the demand for the firm's products will be zero.

Outcaste Labor. No firm can bid a wage higher than θ_{sc} for outcaste labor and receive a profit if this bid is accepted. For a firm to pay a higher wage than θ_{sc}, it must employ outcaste labor in skilled or unskilled jobs. Its profits per laborer will not exceed

$$p\theta_{sk} - \theta_{sc}.$$

If profits are nonnegative,

$$p \geqq \theta_{sc}/\theta_{sk}.$$

But at a price as great as θ_{sc}/θ_{sk} the firm will have no customers: any prospective customer will be outcasted and expect to receive a wage θ_{sc}. Consider this customer. He will buy α units from this firm at a price p and will purchase $(\theta_{sc} - \alpha p)$ units of other goods from other firms. Therefore, his utility will be no greater than

$$\theta_{sc} - \alpha\theta_{sc}/\theta_{sk} + \alpha. \tag{6}$$

But since (6) is less than θ_u by (3), this firm will have no customers. Hence the maximum bid for outcaste labor will be θ_{sc}.

Equilibrium-Breaking Coalition

Finally, a coalition of k^* firms, $k^* > (\theta_u - \theta_{sc})/\alpha(1 - \theta_{sc}/\theta_{sk})$ can break the equilibrium. Such firms can offer a wage bid θ_{sc} for outcaste labor, and offer to sell their output at a price θ_{sc}/θ_{sk}. The expected utility of a person purchasing from these firms will be

$$\min\,(\theta_{sk},\ \theta_{sc} - k^*\alpha\theta_{sc}/\theta_{sk} + k^*\alpha),$$

which is greater than θ_u if $k^* > (\theta_u - \theta_{sc})/\alpha(1 - \theta_{sc}/\theta_{sk})$. Thus the coalition of firms will be able to attract customers; and since workers will be better off receiving θ_{sc} in wages and purchasing from firms that break the caste code, these firms will also be able to attract workers.

Comments on Caste Equilibrium

1. The equilibrium described has two types of distortions due to caste structure. The equilibrium is not Pareto optimal, since in a Pareto-optimal equilibrium N workers would work in skilled jobs, for which they are fully qualified. Also, income distribution is skewed along caste lines, since in the absence of caste all workers would receive the same wage.
2. There is another equilibrium, also with fulfilled expectations, in which all workers work in skilled jobs and receive a wage θ_{sk}. The price of all goods is 1.

A.3 Full Survey Questions

- **General demographic and experience questions**
  - How familiar were you with the content of the paper prior to the reading? (5-point scale)
  - Have you previously read this excerpt prior to today? (Yes/No)
  - I had completed the reading in the allotted time. (7-point Likert scale)
  - How easy or difficult did you find this excerpt? (7-point scale)
  - What percentage of the paper do you think you understand? (Sliding scale from 0 to 100)
- **Two summary questions about the passage**
  1. What is the main argument of the paper?
  2. How does Akerlof support this argument? Please be as specific as possible in your answer.
- **Five specific questions about the passage**
  - **"The Market for Lemons"**
    1. In the excerpt, Akerlof focuses on two groups of traders. Briefly describe their difference and why this matters to the argument of the paper.
    2. Akerlof won the Nobel Prize in 2001 for the paper from which this excerpt was taken. What do you think this paper contributed (to the field of Economics, or otherwise) to merit this award?

3. In the section following the excerpt, Akerlof writes, “It is a well-known fact that people over 65 have great difficulty in buying medical insurance. The natural question arises: why doesn’t the price rise to match the risk?” Based off of the excerpt you read, what do you think was his answer to this question?

4.

$$D(p) = (Y_2 + Y_1)/p \qquad p < 1$$

$$D(p) = (Y_2/p) \qquad 1 < p < 3/2$$

$$D(p) = 0 \qquad p > 3/2$$

These are a set of equations you encountered in the excerpt. Please interpret, to the best of your ability, what they mean and their importance & relationship to the larger position of the excerpt.

* What does p refer to in these equations?
* What do the Ys refer to in these equations?

5.

$$U_2 = M + \sum_{i=1}^{N} 3/2x_i$$

This is another equation you encountered in the excerpt. Please interpret, to the best of your ability, what it means and its importance & relationship to the larger position of the excerpt.

* What does M refer to in the equation?
* What does x refer to in the equation?

* What is the summation over in the equation?

– **"The Economics of Caste and of the Rat Race"**

1. In the excerpt, Akerlof focuses on three groups of people. Briefly describe how they are formed, their difference, and why this matters to the argument of the paper.

2. In the latter half of the excerpt, Akerlof focuses on the idea of caste-class equilibrium. What is the greater purpose of this section, and how can the principles he laid out be applied to another situation outside of the Caste system?

3. Akerlof writes in the introduction to this section the following:
"In Japan as merchants have become more economically successful, so too have the taboos against trade and manufacture been reduced [...] The best example of economic success reducing taboos is most probably, the elimination of the sanctions against collection of interest. The usurer of the Middle Ages has turned into the banker of today."
Why did Akerlof include these examples in the introduction, and how is it related to the excerpt?

4. $$k* > (\theta_u, \theta_{sc})/\alpha(1 - \theta_{sc}/\theta_{sk})$$

   This an equation you encountered in the excerpt. Please interpret, to the best of your ability, what it means (with as much specificity of variables as possible) and its importance & relationship to the larger position of the excerpt.

* What does k* refer to in the inequality?
* What does the right side of the inequality represent? How (generally) did Akerlof reach this expression?

5. The following is a quote from the excerpt: "No firm can bid a wage higher than θ_{sc} for outcaste labor and receive a profit if this bid is accepted. For a firm to pay a higher wage than θ_{sc}, it must employ outcaste labor in skilled or unskilled jobs."
   * What steps did Akerlof undergo to show the claim? (In other words, how did he build up to this claim?) Please be as specific as possible in your answer.

A.4 Amazon Mechanical Turk Grading Rubric

Below is the rubric for each question given to workers on Amazon Mechanical Turk to grade. This was first developed conceptually and then refined using empirical responses as a loose guide for expected detail of answers.

Market for Lemons

What is the main argument of the paper?

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	n/a
1	[wrote something]
2	information symmetry,' 'information asymmetry' [or] market, buyers, sellers
3	symmetry is good [or] asymmetry is bad
4	(a)symmetry **affects** quality of goods (/cars/lemons) [or] quality of information
5	Asymmetry results in a market where no **trades occur**

How does Akerlof support this argument?
Please be as specific as possible in your answer.

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	n/a or no other content
1	[wrote something]
2	Divide into symmetric and asymmetric groups/2 groups
3	Shows that the symmetric group is better off [or] asymmetric group is worse off
4	Compare utility functions [or] compare demand
5	Different **information / probability distributions / quality uncertainty** results in different utility/demand [or] has most of the previous levels and **has included additional steps** outside of the previous levels

In the excerpt, Akerlof focuses on two groups of traders. Briefly describe their difference and why this matters to the argument of the paper.

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	n/a [or] I don't know [or] no other content
1	[Wrote something]
2	Different utilities [with no coefficients mentioned] [or] one group has cars and the other does not
3	The groups are buyer/seller of cars [or] gives specific utility coefficients (1 and 3/2)
4	[The above] + this **matters** because it causes difference in demand in the market between the two groups [or] this **matters** because it shows difference in utility gained from trades
5	[The above] + this **matters** because it causes difference in demand that **results in** (**no/zero**) trades occurring

Akerlof won the Nobel Prize in 2001 for the paper from which this excerpt was taken. What do you think this paper contributed (to the field of Economics, or otherwise) to merit this award?

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	n/a [or] I don't know [or] no other content
1	[Wrote something]
2	**Affects** markets/market phenomenon
3	Explains market **inefficiency**/failures [or] why demand is driven to **zero**
4	Explains market inefficiency/failures **resulting from imperfect information** or information **asymmetry**
5	We can **improve** inefficiency by regulating information [or] [response relates to **why this can help current markets**]

In the section following the excerpt, Akerlof writes, "It is a well-known fact that people over 65 have great difficulty in buying medical insurance. The natural question arises: why doesn't the price rise to match the risk?" Based off of the excerpt you read, what do you think was his answer to this question?

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	n/a [or] I don't know [or] no other content
1	[Wrote something] [responses that include 'demand' or 'equilibrium']
2	Elderly do not know or trust the **quality of insurance** provided
3	**Asymmetric** information [AND] mentions raising price
4	Insurance providers **do not know** how healthy the elderly is
5	Price would be too high for the quality [OR] price cannot **match the risk**

These are a set of equations you encountered in the excerpt. Please interpret, to the best of your ability, what they mean and their importance & relationship to the larger position of the excerpt.

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	n/a [or] I don't know [or] no other content
1	Total demand curves, [but **also** incorrectly **mentions**] **asymmetry**
2	Total demand curves [**does not mention asymmetry**]
3	Explains what each case means (curve 1 = both have demand, curve 2 = only second group has demand, curve 3 = no demand) [or] [**mentions**] **symmetry**

What does p refer to in these equations?

Points	If response includes
0	[Anything else]
1	Price

What do the Ys refer to in these equations?

Points	If response includes
0	[Anything else]
1	Income

This is another equation you encountered in the excerpt. Please interpret, to the best of your ability, what it means and its importance & relationship to the larger position of the excerpt.

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	n/a [or] I don’t know [or] no other content
1	Utility
1.5	Utility of **group 2**
2	[ALL OF THE ABOVE AND] compares with **group 1** [OR] includes the words “supply” or “demand” or “equilibrium” of “market”

What does M refer to in the equation?

Points	If response includes
0	[Anything else]
1	Consumption/utility of goods [other than automobiles]

What does x refer to in the equation?

Points	If response includes
0	[Anything else]
0.5	**Utility** of car
1	**Quality** of the [i'th] car

What is the summation over in the equation?

Points	If response includes
0	[Anything else]
1	All **cars/automobiles**

Economics of the Caste System

What is the main argument of the paper?

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	N/A or nothing relevant written
1	[wrote something relevant]
2	Caste
3	Caste model/system
4	Under certain circumstances, people can break out of the caste system
5	Caste system rules make it so that people stay within it/caste equilibrium

How does Akerlof support this argument?
Please be as specific as possible in your answer.

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	n/a or no other content
1	[wrote something]
2	Includes description of 3 types of laborers, market structure
3	Gives generic structure of proof (structure, formal model, claims)
4	Gives structure of proof with specific details
5	Arrives at a conclusion: Firm trying to pay outside of the caste system will lose profit

In the excerpt, Akerlof focuses on three groups of people. Briefly describe how they are formed, their difference, and why this matters to the argument of the paper.

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	n/a [or] I don't know [or] no other content
1	[Wrote something]
2	Skilled, unskilled, scavenger / N, D, outcasts
3	Explains difference between the castes
4	Explains how they're formed (Response is related to production value)
5	Explains why it matters to the paper (Response is related to supporting mathematical model of the caste system/proof in equilibrium)

In the latter half of the excerpt, Akerlof focuses on the idea of caste-class equilibrium. What is the greater purpose of this section, and how can the principles he laid out be applied to another situation outside of the Caste system?

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	n/a [or] I don't know [or] no other content
1	[Wrote something]
2	Explains purpose of this section (vague, mentions equilibrium)
3	Explains purpose of this section (more specific, related to individual unable to disturb the balance)
4	Able to relate it to external case (but related to social structures)
5	Able to relate it to external case (creative, not related to simple social hierarchy [or] very specific and in-depth)

Akerlof writes in the introduction to this section the following:
"In Japan as merchants have become more economically successful, so too have the taboos against trade and manufacture been reduced [...] The best example of economic success reducing taboos is most probably, the elimination of the sanctions against collection of interest. The usurer of the Middle Ages has turned into the banker of today."
Why did Akerlof include these examples in the introduction, and how is it related to the excerpt?

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	n/a [or] I don't know [or] no other content
1	[Wrote something]
2	Vaguely mentions caste and economics
3	Compares groups in the quote to outcastes
4	Mentions economic methods of modifying/changing caste system
5	People in lower levels are able to change their position through monetary means / going against the norm [or] more specificity on changing the caste system in relation to the low levels

This an equation you encountered in the excerpt. Please interpret, to the best of your ability, what it means (with as much specificity of variables as possible) and its importance & relationship to the larger position of the excerpt.

Points	If response includes… (cumulative; each point includes most information from previous levels as well)
0	n/a [or] I don't know [or] no other content
1	Mentions theta/definition of variable
2	Cannot break caste equilibrium
3	Breaking caste equilibrium

What does k* refer to in the inequality?

Points	If response includes
0	[Anything else]
1	Coalition of firms

What does the right side of the inequality represent? How (generally) did Akerlof reach this expression?

Points	If response includes
0	[Anything else]
1	Ratio of labor

The following is a quote from the excerpt: "No firm can bid a wage higher than θsc for outcaste labor and receive a profit if this bid is accepted. For a firm to pay a higher wage than θsc, it must employ outcaste labor in skilled or unskilled jobs."

What steps did Akerlof undergo to show the claim? (In other words, how did he build up to this claim?) Please be as specific as possible in your answer.

Add Points	If response includes…
0	n/a [or] I don't know [or] no other content
1	[wrote something relevant]
1	Mentions firm constraints
1	Firm is profit-maximizing / receives zero profits / nobody would purchase
1	Mentions scenario where firm pays for outcaste labor or higher for another caste
1	Mentions negative consequences of a deviation from equilibrium

References

[1] Acar, A. & İşisağ, K. U. (2017). Readability and Comprehensibility in Translation Using Reading Ease and Grade Indices. *International Journal of Comparative Literature and Translation Studies*, 5(2), 47–53.

[2] Cowan, N. (2008). Chapter 20 What are the differences between long-term, short-term, and working memory? In W. S. Sossin, J.-C. Lacaille, V. F. Castellucci, & S. Belleville (Eds.), *Progress in Brain Research*, volume 169 of *Essence of Memory* (pp. 323–338). Elsevier.

[3] Danili, E. & Reid, N. (2006). Cognitive factors that can potentially affect pupils' test performance. *Chem. Educ. Res. Pract.*, 7(2), 64–83.

[4] Hindal, H., Reid, N., & Badgaish, M. (2009). Working memory, performance and learner characteristics. *Research in Science & Technological Education*, 27(2), 187–204.

[5] Jung, E. S. & Reid, N. (2009). Working memory and attitudes. *Research in Science & Technological Education*, 27(2), 205–223.

[6] Peng, P., Namkung, J., Barnes, M., & Sun, C. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology*, 108(4), 455–473.

[7] Tang, R., Shen, B., Sang, Z., Song, A., & Goodale, M. A. (2018). Fitts' Law is modulated by movement history. *Psychonomic Bulletin & Review*, 25(5), 1833–1839.

[8] Wertheimer, M. (2012). On perceived motion and figural organization.