Modeling Student Comprehension
Using Textbook Annotations:
An Exploration of a Large Scale, Naturalistic Corpus

by

David Y.J. Kim

B.A., Sogang University, 2016

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Master of Science
Department of Computer Science
2021

Committee Members:
Michael C. Mozer Ph.D., Chair
James Martin Ph.D.
Sidney D’Mello Ph.D.
Kim, David Y.J. (M.S., Computer Science)

Modeling Student Comprehension
Using Textbook Annotations:
An Exploration of a Large Scale, Naturalistic Corpus

Thesis directed by Prof. Michael C. Mozer Ph.D.

Education is slowly moving toward integrating digital textbooks. Due to the ability of collecting students engagement data in digital textbook settings, we see valuable opportunities to observe students learning process. Student’s interaction with textbooks may serve as a window into their mental state, level of comprehension, grasp of the key idea, and so on. In this study, we are interested in investigating whether highlights can work as a data source in modeling student comprehension. We explore this hypothesis via observational data in a genuine educational setting with the aid of OpenStax Tutor learning platform, a non-profit organization that supports open-access college-level digital textbooks. During the assignments, students were able to highlight and add annotations to their e-textbooks while reading. Comprehension of the material was assessed by a quiz that students take at the end of each section. We aim to predict quiz performance in an authentic digital learning environment and determine the effectiveness of models that leverage different representations of highlights such as the pattern of highlights or content of the highlights a student produces. We find that when students choose to highlight, the specific pattern of highlights can reliably explain the variance in observed quiz scores (7% ~ 13%). We explore many different representations of the pattern of highlights and discover that a low-dimensional logistic principal component-based vector is most effective as input to a ridge regression model. We further investigate using the content of the highlights. Using the semantic representation of the highlighting, we built probabilistic models to predict quiz performance. Our results show that quiz score prediction accuracy reliably improves with the inclusion of highlighting data (by about 1% ~ 3%) over the baseline model. For different held-out settings, we found positive results
both for held-out students and for held-out student questions (i.e., randomly holding out student, question pairs), but not for held-out questions. Furthermore, we found these results consistent in both factual questions and inference level questions.
Dedication

This thesis is dedicated to my father, the man who taught me to perform all of life’s tasks, no matter how big or small, to the best of my ability and without complaint. He’s the man I will always aspire to be.
Acknowledgements

This research is supported by NSF awards DRL-1631428 and DRL-1631556. We thank Tyler R. Scott for their suggestions on this research and for helpful feedback on earlier drafts of this manuscript.
## Contents

Chapter

1 Introduction 1

1.1 Research Overview 1

1.2 Paper Outline 4

2 Prior Research and Background 5

2.1 The Role of Highlighting in Education 5

2.2 Limitations of Prior Studies & Necessity for Further Investigation 7

2.3 Highlighting statistics and student performance 10

3 Analysis using the positional highlight representation 14

3.1 Data 14

3.1.1 Openstax data explanation 14

3.1.2 Data 17

3.2 Methodology 19

3.3 Results 22

3.3.1 Replication in other data 24

4 Analysis using the Semantic highlight representation 27

4.1 Data 28

4.2 Methodology 30
4.2.1 Background on BERT                                      30
4.2.2 Representing the semantic similarity between highlights and quiz questions 32
4.2.3 Development of the augmented Rasch model                  35
4.3 Results                                                  37
  4.3.1 Results based on different hold-out settings            38
  4.3.2 Model prediction accuracy based on Bloom taxonomy       41
  4.3.3 Model prediction accuracy with syntactic versus semantic highlighting representations 43

5 Conclusion and Further Research                              45

Bibliography                                                 48
### Tables

**Table**

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Between student comparison</td>
<td>11</td>
</tr>
<tr>
<td>2.2 Within student comparison</td>
<td>12</td>
</tr>
<tr>
<td>3.1 Overview of 2018 OpenStax Data</td>
<td>19</td>
</tr>
<tr>
<td>4.1 Overview of 2019 OpenStax Data</td>
<td>29</td>
</tr>
</tbody>
</table>
Figures

Figure

2.1 Mean scores for highlighted versus non-highlighted sections, by course topic. Error bar indicate one standard error of the mean 12

3.1 This figure is originally from “Highlighting Associated with Improved Recall Performance In Digital Learning Environment.”, Waters et al. (2020) 15

3.2 Samples of student highlighting from one section of the biology textbook, the focus of which is on the scientific method. The student whose highlights are shown in the upper portion of the figure scored 54% on the section quiz; the student whose highlights are below scored 100% 18

3.3 Sample material from one section of the biology textbook, the focus of which is on the scientific method. The intensity of the red (blue) color indicates a model’s prediction of increased (decreased) accuracy on the quiz when the corresponding word is highlighted. This result comes from the model that used the PCA(10%) representation of highlights, and for this section, this model used only the first principal component 21

3.4 Proportion of variance ($\rho^2$) in quiz scores explained by a student’s specific highlighting pattern. The bars show one standard error of the mean. Each point along the abscissa corresponds to building a model with a particular representation of the highlighting pattern, as described in the text 23
3.5 Proportion of variance ($\rho^2$) in quiz scores explained by a student’s specific highlighting pattern. The bars show one standard error of the mean. Each point along the abscissa corresponds to building a model with a particular representation of the highlighting pattern, as described in the text.

4.1 Left Passage is an example of section (The Science of Biology), while the query on the top right is a question asking about a concept of the section. The framework on the bottom side is how we intend to find the similarity score between the highlighted sentence and the question.

4.2 Question is from the section “The Science of Biology” and the following five sentences are the semantically most similar sentences related to the question.

4.3 The graph above is a simulation of random $10^4$ numbers from a uniform distribution $U(0, 2)$ and calculating the $norm_p$ value.

4.4 Each bar indicates the average of these metrics across all sections and error bars reflect $+/-1$ SEM, corrected to remove variance due to the random factor [44]. Each point along the abscissa corresponds to combination of different features. The lighter color bars indicate performance when the features are used alone, the darker color bars indicate when the highlight features are added.

4.5 Each bar indicates the average of these metrics across all sections and error bars reflect $+/-1$ SEM, corrected to remove variance due to the random factor [44]. Each point along the abscissa corresponds to combination of different features. The lighter color bars indicate performance when the features are used alone, the darker color bars indicate when the highlight features are added.
4.6 Each bar indicates the average of these metrics across all sections and error bars reflect \(+/-\) 1 SEM, corrected to remove variance due to the random factor \([44]\).

Each point along the abscissa corresponds to combination of different features. The lighter color bars indicate performance when the features are used alone, the darker color bars indicate when the highlight features are added.

4.7 The number between the parenthesis indicate the Bloom’s Taxonomy level of questions that were predicted. The lighter color bars indicate performance of the Rasch Model, the darker color bars indicate when the highlight features are added to the IRT model. Error bars reflect \(+/-\) 1 SEM, corrected to remove variance due to the random factor \([44]\).

4.8 Each point along the abscissa corresponds to different models. Syntactic IRT model indicates using ability, difficulty, and the first component of PCA vector for the model features. Semantic IRT model indicates using ability, difficulty, and the semantic embedding for the model features. Error bars reflect \(+/-\) 1 SEM, corrected to remove variance due to the random factor \([44]\).
Chapter 1

Introduction

1.1 Research Overview

Digital textbooks have become increasingly available with the popularity of e-readers, the advent of open-access learning resources such as Openstax, and reduced cost of these new platforms. Like other researchers in AI and education, we see valuable opportunities to observing students as they interact with their textbooks and become familiar with new material. For mathematics or physics courses, where students can demonstrate their understanding by working through exercises, researchers have long had the opportunity to observe students’ problem-solving skills and to suggest hints and guidance to remediate knowledge gaps [58]. However, in courses where textbooks contain factual material, such as in biology or history, opportunities for inferring student understanding are more limited. The obvious means is quizzing a student after they have read a section of text, but quizzes are unpleasant and time-consuming to students, who often fail to appreciate the value of such quizzes to bolstering long-term knowledge retention. Consequently, we have been investigating implicit measures we can collect as students interact with a digital textbook, measures which do not require students to explicitly demonstrate their understanding, as is required in a quiz. To give some compelling examples of implicit measures, an active study is being conducted whether students eye gaze could be used to predict mind wandering\(^1\) during reading [46, 25]. Other studies collected studying behavior from video streams to predict mind wondering [10]. While a finished

---

\(^1\) Attentional shift away from the processing of task-related information to the processing of task-irrelevant thoughts or ideas
In this thesis, we are interested in highlighting—the yellow marks and underlines that students make in a textbook in order to emphasize the material that they perceive as particularly pertinent. Highlighting is a popular study strategy \cite{47} and students voluntarily highlight because they believe it confers learning benefits \cite{22}. Given that highlights reflect material that students believe to be important, one has reason to hypothesize that highlights could be useful for assessing comprehension. Although there is well-established research literature in educational psychology examining whether highlighting benefits student learning \cite{45, 14, 50, 39}, our focus is on the question of whether highlights can be used as a data source to predict student comprehension and retention. The advantage of this data source is that it imposes no burden on students. Also, most of these research were based on paper text annotation and highlighting, but the impact of note taking and highlighting behaviors on reading comprehension may differ based on format of instruction (online, paper etc.). Kobrin and Young \cite{37} propose some characteristics of digital display that could interfere one’s cognitive processes required for comprehension of the text. These characteristics include visual fatigue, inability to directly mark text, and the inability to see the entire text at once. This interference of comprehension can also affect working memory \cite{5}, which is a system for holding information and allowing it to used to perform a wide range of cognitive tasks. This working memory system is known to serve as the basis for conceptual comprehension and provide capacity for complex thoughts \cite{53}. With the premise that working memory is true, studying would be most efficient when working memory is uninterrupted. Empirical evidence suggests that reading from a digital screen has negative effects on comprehension due to visual fatigue and focal distractions \cite{43, 31}. As a result, readers may have to expend mental resources on decoding information presented on a digital screen rather than using that cognitive capacity for comprehension. Due to the reasons that are mentioned above, studying strategies deemed to be effective or ineffective in paper-textbooks might not conveniently translate in digital settings, thus as educators we are interested whether previous well known study strategies work in digital settings.
We focused on identifying the relationship between the pattern of highlighting and subsequent retention and understanding. While exploring alternative approaches to representing the pattern of highlights, the fundamental assumption we adopted is that the feature representation should be compact to reduce redundancy but still expressive enough to capture the signal in the data. Our initial approach was to parse the section based on sentences, words, or any other different standards and project them to a high dimensional vector where each cell indicates whether the corresponding chunk is highlighted or not. We use the term **positional analysis** for this approach. This approach included investigations of different methods for parsing the text—including at the word level, sentence level, and parsing into fixed number of characters per chunk (e.g., 100, 200, ..., 1500). Because this representation tends to be very high dimensional we also attempt to use principal component analysis to reduce redundancy. Based on our investigation, we have found out that parsing the section into words then reducing the dimension using principal component analysis and picking only the top 10% explanatory variables explains the most variance of quiz scores.

A complementary approach we investigate in this thesis is to consider a **semantic analysis** of the highlights. Using BERT [21], it is possible to determine the semantic similarity of two sentences. In particular, we focus on the similarity of quiz questions to both highlighted and non-highlighted sentences. We investigated if this similarity score can be used to improve the prediction of quiz performance on specific questions. One problem we faced is that students highlight different numbers of sentences and thus it is necessary to aggregate similarity scores into a fixed-size feature vector. We find that quiz score prediction accuracy reliably improves with the inclusion of highlighting features. Also, the semantic representation is superior to the syntactic representation. Finally, including highlighting features improves model predictions for questions at all levels of the Bloom hierarchy [9], giving us hope that better highlight representation might be useful in measuring student’s deep conceptual comprehension.
1.2 Paper Outline

Ongoing forward in Chapter 2 we will overview the past literature of highlighting, we also add a simple statistical analysis regarding whether highlighting associates with higher quiz scores. Finally, we point out the critical drawbacks of previous studies and justify our research. In Chapter 3 we introduce our first approach the **positional** analysis. Beforehand we will explain how the Openstax Tutor works and collect data, following will be an overview of the metadata. We will explain the methodology of the syntactic analysis and its results. Chapter 4 will introduce our next approach the **semantic** analysis. Since we used a new set of data for this analysis, we briefly go over the metadata of the new data. We move on to explaining the methodology of the semantic analysis and its results. Chapter 5 will wrap up the whole thing by summarizing the research and discussing issues that should be addressed in future research.
Chapter 2

Prior Research and Background

2.1 The Role of Highlighting in Education

Highlighting is a common studying strategy among students [2, 3, 29] and given that highlights reflect material that students believe to be important, one has reason to hypothesize that highlights could be useful for assessing comprehension. There exists a long history of research concerning the efficiency of underlining/highlighting\(^1\) in terms of studying [45]. Despite the abundant research, definitive conclusions have yet to be reached whether highlighting is truly an efficient study strategy due to mixed results. Some studies have found that highlighting is an efficient strategy [14]. However, most have concluded that the effect of highlights is negligible [49, 45, 23, 4, 50]. In fact, one study claimed that highlight provides "low utility" for students [22]. One interesting study that attempts to explain the possible reason for the above results is that low-skill readers report more reliance on highlighting strategies and actually mark their text more than better readers [6]. While low-skill readers rely on highlighting, they struggle to find the most relevant material to highlight, leading to a lack of efficacy.

Based on previous literature, highlighting might not be universally beneficial in one’s study strategy, but we might consider some contextual nuances of student highlights. Highlights may function by directing some form of enhanced cognitive processing to the highlighted information [24]. A recent study suggests two ways of possibility how it might enhance cognitive processing [62]. It may encourage a deeper level of processing of the highlighted information [18], or it may

---

\(^1\) Note that previous studies have shown that there is no difference between underlining and highlighting [39]
function as an external memory [27]. In support of this view, studies have found that students were more likely to answer questions correctly when the question is related to the information they highlighted, compared to questions that have no related highlights [14, 39, 33, 35]. There is also some evidence that producing a highlight is more beneficial when the student themselves actively highlight the material rather than when passively reading highlights produced by someone else [50].

If it is the case that highlighting enhances learning only to the information that is highlighted, then the specific information that is highlighted matters considerably. This conclusion was reached by one previous study [39]. They explored the effects of highlights asking undergraduates to read articles from Scientific American in four highlighting conditions: active, in which students highlighted as much content as they wanted to; passive yoked, in which students read marked texts that had been highlighted by yoked participants in the active condition; passive expert-based, in which students read marked texts that had been highlighted by the experimenters to reflect critical material; and control, in which students read articles without any highlights. Based on their findings, there was no significant difference between conditions in overall exam scores. Nonetheless, they observed if the sentences in the text that is critical for a given question had been highlighted, that specific question performance benefits significantly. This phenomenon was more apparent when active highlighters were the ones that highlighted the segment compared to passive-yoked highlighters. Also, passive expert-based highlighting yields superior performance on these questions compared to the control condition. These findings provide weak yet possibilities that highlighting engages in a deeper level of the learning process. On the other hand, because highlights influence readers to focus mainly on the highlighted material, some view that highlights facilitate the von Restorff effect [39, 49, 14, 12, 61]. This view is supported by one study, where they found that students who highlighted and later reviewed their annotation performed significantly worse than non-highlighters on higher-level inference tasks [52]. However, this must be further investigated as not much study has been conducted about the effects of highlights across multiple levels of

---

2 The Von Restorff effect predicts that when multiple homogeneous stimuli are presented, the stimulus that differs from the rest is more likely to be remembered.
knowledge.

2.2 Limitations of Prior Studies & Necessity for Further Investigation

In this section, we note limitations to the existing literature on highlighting. One limitation is that most previous literature results relies on laboratory findings [62, 18, 22, 39]. Laboratory experiments have the advantage of collecting data in controlled settings, but one might question whether laboratory results can be applied to real-world settings involving actual course content in an authentic learning environment. One reason why laboratory setting might be problematic is that most laboratory experiments are conducted in a timed manner. This could make the participants feel rushed and not have enough time to digest the passage enough to realize what the main point is, hence, highlighting ineffectively. One study [62] increased their time of reading because participants showed concerned about the rushed reading time. While assignments of academic courses still have deadlines, they are usually ranged for days and weeks which is more than enough for students to fully digest the material and highlight after understanding the full picture. Another concern is that participants in laboratory experiments might as not be fully incentivized to perform well in the experiment. Studies have provided bonuses [62] in order to incentive participants, and in fact, one study found that there is a difference in performance when these bonuses are present or not [26]. In contrast, students in academic course are studying for their grades which will naturally enhance motivation. Therefore, it is reasonable to think that the effects of highlighting may differ between populations of laboratory participants and real students.

Another limitation is the small sample sizes that existing highlighting literature have experiment/analyzed with. Ranging from tens to at most a few hundreds of participants, 150 is known to be the average sample size for highlighting research [24]. Considering the many factors that influences one’s quiz performance (e.g., the motivation for the task, the subject of the text) it is not surprising that previous studies had a hard time to find the small signal that highlighting provides with such small sample. In our study we have collected samples from thousands of students throughout 4 semesters and we show that this high number of data will be crucial in obtaining
more reliable and generalized results.

The third limitation is that most studies on highlighting were based on the traditional paper text using pens or highlighters to mark. As majority of previous highlighting studies were conducted prior to 1995 [24] when reading textbooks with a computer was not common, it is natural of them to be based on paper text. However, currently more students are starting to use tablets or their computer to read textbooks and many educators are also aiming to build intelligent digital textbooks that would achieve active learning [1]. We cannot assume that results obtained based on paper material will apply in digital settings with no difference. This statement is reinforced with some recent studies that exists a fundamental differences between learning with digital resources and paper ones. A recent meta analysis research found that Paper-based reading yields better comprehension outcomes than digital-based reading and the advantage of paper-based comprehension has increased over the years since the year 2000 [20]. Other studies found that students notes differ when using pen/paper compared to digital notes [48] and overall studying habits change while studying in digital settings [19]. Moreover, while highlighting is a common practice while reading paper media, it seems to be far less common when participants read digital media [41].

Finally, we wanted to point out that most of the past research used recall level questions, questions that require remembering the definition of a word or simply reiterating a phrase in the text, as their measurements or only found positive results in recall questions [24]. Even though predicting student’s recall performance is a valuable source for educators, most educators strive to teach students higher-level questions such as inference or application. There are two possibilities we hypothesize why studies until now struggled to find highlighting effects on higher-level questions. One is that simply the act of highlights does not affect the cognitive process deep enough to reach higher-level concepts. Another is the highlight representation used in previous literature was not powerful enough to capture the subtle information that highlighting acts provide. Most previous research used simple representations of highlights such as, ”Did the student highlighted or not” [18, 50, 22, 39] or ”Did the student highlight the relevant phrase?” [24]. It was only recently, studies have started to investigate different representations of highlights such as highlighting pattern
and we also attempted new representations. In our study, we investigated whether our new highlighting representation can find efficiency in inference level questions, a popular way of characterizing levels of learning is Bloom’s Taxonomy [9, 38, 15], which organizes knowledge into six levels - recall, understand, apply, synthesize, evaluate, and create. Only a few studies to our knowledge have examined the influence of highlighting on cognitive levels higher than recall [24, 7, 35]. It is important to understand how the act of highlighting affects performance on different levels of cognitive processes.

Recently, our team conducted two studies that provide preliminary evidence in support of the hypothesis that highlights provide insight into comprehension. In a laboratory experiment, Winchell et al. [62] asked participants to read and optionally highlight three sections of an OpenStax biology textbook [16], chosen with the expectation that the passages could be understood by a college-aged reader with no background in biology. The three passages concern the topic of sterilization: one serving as an introduction, one discussing procedures, and the last summarizing commercial uses. Participants were told that they would be given a brief opportunity to review each of the passages and the highlights they had made, and would then be quizzed on all three passages. The quiz consisted of factual questions concerning the material, both in a multiple choice and fill-in-the-blank format. The purpose of the limited-time review was to incentivize participants to highlight material to restudy during the review phase. Winchell et al. find reliable improvements in the accuracy of predicting correctness on individual quiz questions with the inclusion of highlighting patterns, both for held-out students and for held-out student-questions (i.e., questions selected randomly for each student), but not for held-out questions. However, the accuracy of predicting the correctness of a student’s answer increases by only 1-2%.

In contrast, Waters et al. [24] explored the impact of highlighting produced by real students enrolled in actual college-level courses in Biology, Physics, and Sociology. Students read textbooks and highlighted as they wished on a digital learning platform, OpenStax Tutor. At the end of a section, they answered three practice questions before moving on to the next section. The data set included 4,851 students, 1,307 text sections, and a total of 85,505 student highlights. Waters et al.
found an effect of highlighting on learning outcomes: for questions tagged as “recall” on Bloom’s taxonomy scale, a small but reliable increase in a student’s accuracy on a particular question is observed if the student highlights the critical sentence in the text needed to answer the question.

Neither of these studies is completely satisfying. The Winchell et al. study was conducted via Mechanical Turk with 200 participants with unknown motivation levels. It involved just three passages and twelve quiz questions (formulated either as multiple choice or fill-in-the-blank) and took place over 40 minutes. Consequently, its application to authentic digital learning environments is unclear. The Waters et al. study was on a much larger scale in the context of actual coursework, but their predictive models were limited in scope: the models considered only the highlighting of a critical sentence, whereas Winchell et al. constructed predictive models based on the pattern of highlights in the section. It’s possible that the strongest predictor of subsequent recall by a student may not be whether the critical sentence was highlighted but by the highlighting of material the precedes or follows the critical sentence.

2.3 Highlighting statistics and student performance

In prior research, when specific highlights of individual students are concerned, they are typically regarded in terms of some statistics variable such as the number of key points highlighted, the efficiency of the highlighted sentences, the proportion of sentences highlighted.

We investigated in a digital setting whether highlighting is associated with higher quiz scores. We first report on some baseline analyses showing that the choice to highlight at least some material is associated with higher quiz scores. Although we cannot ascertain a causal relationship, we can eliminate some confounders. Waters et al. [24] conducted a similar analysis via a latent-variable model that included highlighting as a feature. However, they focused on predicting the correctness of response to a particular question based on whether or not the corresponding critical sentence in the text had been highlighted. We performed a more fundamental investigation of the relationship between mean accuracy across all questions and whether or not the student had highlighted any material.
In a first analysis, we divided sessions by course topic and by whether students had made highlights during that session. In Figure 2.1a, we show mean scores with ±1 SEM bar by course topic. Across all topics, highlighted sessions are associated with higher quiz scores than non-highlighted sessions (Table 2).

This analysis of course does not indicate that highlighting has a causal effect on performance. Possible non-causal explanations include:

- More diligent students may tend to highlight and more diligent students study hard and therefore perform better on quizzes.
- Whether or not a student highlights may be correlated with the difficulty of the material.

For example, a student who is struggling to understand material may not feel confident to highlight, leading to lower scores for non-highlighted sections.

To address these explanations, we conducted further analyses. To rule out differences in student diligence being responsible for the effect, we performed a within-student comparison of highlighted verses non-highlighted sections. We consider only students who have both highlighted and non-highlighted sections, and we compute the mean score by the student when they highlight and when they do not highlight. This within-student comparison is shown in Figure 2.1b. For each of the three-course topics as well as a mean across topics. We find the same pattern as before that mean student scores are higher for highlighted than for non-highlighted sections (Table 3). However, the difference for Sociology is not statistically reliable.

To rule out the possibility that the decision to highlight is in some way contingent on the difficulty of the section, we used item-response theory, specifically the Rasch model [56], to infer

<table>
<thead>
<tr>
<th>Topic</th>
<th># Highlighted Sessions</th>
<th># Non-highlighted Sessions</th>
<th>t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td>1,998</td>
<td>32,407</td>
<td>t(34403) = 9.38, p &lt; 0.01</td>
</tr>
<tr>
<td>Physics</td>
<td>724</td>
<td>40,283</td>
<td>t(41005) = 6.60, p &lt; 0.01</td>
</tr>
<tr>
<td>Sociology</td>
<td>117</td>
<td>7,405</td>
<td>t(7520) = 2.06, p = 0.04</td>
</tr>
</tbody>
</table>
section difficulty from the student-section observation matrix. We found a minuscule positive correlation of 0.02 between difficulty and the probability of highlighting a section that was not statistically reliable ($p = .61$). We would have expected to observe a negative correlation if the explanation for higher scores with highlighting was due to students choosing to highlight easier material.

Although we haven’t definitely ruled out non-causal explanations for the relationship between highlighting and scores, our results are suggestive that in a digital textbook setting, highlighting may serve to increase engagement which is reflected in improved scores. This finding goes somewhat counter to the research with traditional printed textbooks that fails to find value for highlighting as a study strategy [22].

Our analysis shows that highlighting is correlated with student performance in a digital
setting. The primary goal of the research in this thesis is to ask whether the specific pattern of highlights can provide an indication of the specific knowledge a student will acquire.
Chapter 3

Analysis using the positional highlight representation

In this chapter, we have focused on whether the chunk of phrases students highlight in the passage can be used as a rich data source for inferring student understanding. We conducted our analysis based on data obtained from OpenStax Tutor during two semesters in 2018. Students were enrolled in Biology, Physics, and Sociology courses and read sections of their introductory text as part of required coursework, optionally highlighted the text to flag key material, and then took brief quizzes at the end of each section. After exploring several different representations of the pattern of highlight, our analysis shows that when students choose to highlight, the low-dimensional logistic principal component-based vector can explain about 13% of the variance in observed quiz scores. We obtained new data from OpenStax Tutor for two semesters in the year 2019. The same subjects with the addition of American History. After replicating the analysis, the explanatory power reduced to about 7% but the fact that the low-dimensional logistic principal component-based vector came out as the best representation maintains.

3.1 Data

3.1.1 OpenStax data explanation

As we have continuously mentioned, for this research we have collected data from a web-based learning platform called OpenStax Tutor. OpenStax tutor is both a commercial product and

\footnote{Most of the contents of this chapter were published as Second Workshop on Intelligent Textbooks The 21th International Conference on Artificial Intelligence in Education (AIED’2020) 
}
a browser-based application. It is designed to serve as a companion to several course textbooks, and it is optimized for desktop usage [24]. There are several functionalities and features the OpenStax Tutor platform has, and in this section, we plan to describe some of them. However, not all of the functionality was used as our data and analysis. Thus, this description would be limited to the functionality that is relevant to our analysis.

The “reading assignment” and “homework assignment” functionality of OpenStax Tutor was the center of our study and analysis. The assignments consist of a passage of text that are segments of the textbook and several questions related to the passage that students need to solve. Course instructors can assign assignments by selecting which book section (e.g., Chapter 1-Section 1, Chapter 1-Section 2). As all assignments do, these assignments had a starting date and a due date. Within these dates, students could spend as long as they desired, and could go back to reread a section at any time. After finishing reading and reviewing the sections, students were evaluated about the section with some “core” practice questions. Core practice questions are always related to the book section they had just read.
Aside from the core practice questions, occasionally students will be shown "spaced" practice questions. Spaced-practice questions are questions distributed across sections, they are questions not directly related to what the student has just read but sections they have read previously. Spaced practice questions are designed to improve retention of the materials. The number and timing of spaced practice questions were determined by the instructor. The OpenStax Tutor Platform will automatically choose questions from either one assignment ago or three ago, the number of each varied depending on how many spaced practice questions the instructor included. Openstax tutor prioritized questions that had not been previously seen by students. Figure 3.1 visualize the above process.

All practice questions in Openstax Tutor (core and spaced) were presented in a two-step hybrid format. In the two-step format, students were first required to answer the question with a free-form response. After submitting a free form response, all students were then given multiple-choice options to choose from, regardless of the veracity of their free form response. After making a multiple-choice response, students were immediately shown feedback on the veracity of their multiple-choice response, as well as explanatory feedback about why their response was incorrect (if applicable). While all students with the same instructor would receive the same number of questions, the exact questions given to each student were chosen dynamically. Specifically, Tutor used an algorithm to select questions from a pool of available questions that were an appropriate level of difficulty for the student. While the instructor could not specify which questions were presented for core and spaced practice, they could exclude specific questions they did not like from the pool of available questions in Tutor.

While reading sections, students were able to access an optional highlighting tool to create highlights on the text. To create a highlight, students would use their mouse to select the text they desire to mark with a click-and-drag action. When the student completed their click and drag action, a small pop-up would appear over their selection. When the highlighter icon was selected, the highlight of the selected text would be saved and the highlighted portion would persist with a translucent light blue color. Students were able to delete highlights anytime, and if they keep it
the segment would be highlighted throughout their assignment.

Finally, instructors were not able to add an additional question or edit in their own taste. All questions were written by experts hired by the makers of OpenStax Tutors. All questions were tagged by subject matter experts according to Bloom’s taxonomy.

3.1.2 Data

Our analyses were all conducted by sections (sub-division of the chapters). For each student and each section, we grouped together all questions answered by the student, both core questions and space-practice questions. The student’s score is the mean proportion correct of all questions that are associated with the section. For each score, we recovered the highlighted character positions in the section. These positions consist of the complete list of indices of characters in the section that the student highlighted. The first character in the section is indexed as position 1, and so forth. From the highlighted character positions, one can recover the exact pattern of highlights marked by the student. Because of a quirk in the raw data base, we had to recover the highlighted positions from the unindexed collection of literal words, phrases, and sentences that the student highlighted. In almost all cases, the highlighted positions could be recovered unambiguously. In a few cases, such as when the student highlighted a single word which appeared multiple times in a section, we assumed that the index was its first occurrence in the section. This ambiguity arose very rarely.

We grouped the data by section. Each section is analyzed independently, and we report mean results across sections. Because the textbooks were electronic, they were revised during the time period in which we obtained data. As a result, some sections have multiple versions. We collapsed these revisions together since typically only a few words changed from one version to the next, and it was easy to align the highlighted fragments.

Table 1 presents an overview of the data set. There are a total of 4,851 students, 1,157 distinct sections, and 479,879 sessions, where a session consists of a particular student reading a particular section. Students answered one or more quiz questions in only 328,575 sessions, and
On a global scale, many researchers are committed to finding ways to protect the planet, solve environmental issues, and reduce the effects of climate change. All of these diverse endeavors are related to different facets of the discipline of biology. Escherichia coli (E. coli) bacteria, in this scanning electron micrograph, are normal residents of our digestive tracts that aid in absorbing vitamin K and other nutrients. However, virulent strains are sometimes responsible for disease outbreaks. The Process of Science Biology is a science, but what exactly is science? What does the study of biology share with other scientific disciplines? We can define science (from the Latin scientia, meaning “knowledge”) as knowledge that covers general truths or the operation of general laws, especially when acquired and tested by the scientific method. It becomes clear from this definition that applying scientific method plays a major role in science. The scientific method is a method of research with defined steps that include experiments and careful observation. We will examine scientific method steps in detail later, but one of the most important aspects of this method is the testing of hypotheses by means of repeatable experiments. A hypothesis is a suggested explanation for an event, which one can test. Although using the scientific method is inherent to science, it is inadequate in determining what science is. This is because it is relatively easy to apply the scientific method to disciplines such as physics and chemistry, but when it comes to disciplines like archaeology, psychology, and geology, the scientific method becomes less applicable as repeating experiments becomes more difficult. These areas of study are still sciences, however. Consider archaeology—even though one cannot perform repeatable experiments, hypotheses may still be supported. For instance, an archaeologist can hypothesize that an ancient culture existed based on finding a piece of pottery. He or she could make further hypotheses about various characteristics of this culture, which could be correct or false through continued support or contradictions from other findings. A hypothesis may become a verified theory. A theory is a tested and confirmed explanation for observations or phenomena.

Figure 3.2: Samples of student highlighting from one section of the biology textbook, the focus of which is on the scientific method. The student whose highlights are shown in the upper portion of the figure scored 54% on the section quiz; the student whose highlights are below scored 100%. 
students highlighted portions of the text on only 8,846. One surprising observation is the relative scarcity of highlights during reading, given that students consider highlighting to be a fruitful study strategy. However, highlighting in an electronic text may be awkward or unfamiliar to students.

Nonetheless, we have adequate data to consider with the 8,000+ sessions with highlights. We focus on the sections which had a critical mass of students who highlighted. We identified 28 such sections, with the largest section having 142 highlighters and the smallest section having 31 highlighters. Across the highlighted sessions, the mean quiz score is 75% with a standard deviation 10%.

### 3.2 Methodology

The analysis in the previous chapter simply considered whether or not a student highlighted a section, but ignored a rich information source—the specific words, phrases, and sentences that were highlighted. Our goal is to determine whether positional encoding of highlights help explain scores. Positional encoding of highlights finds regularities such as "if good students tend to highlight sentence 14 but not 28, would it be reasonable to predict someone to be a good student if they highlight sentence 14 but not 28". Here we use only sections of the Biology text, which had the greatest number of student highlighters. We model each section independently and we include only students who highlighted one or more words in the section. Our models predict a specific student’s quiz score from the specific pattern of highlighting that student made. The pattern

---

**Table 3.1: Overview of 2018 OpenStax Data**

<table>
<thead>
<tr>
<th></th>
<th>Biology</th>
<th>Physics</th>
<th>Sociology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
<td>1,946</td>
<td>2,421</td>
<td>484</td>
</tr>
<tr>
<td>Sections</td>
<td>608</td>
<td>435</td>
<td>114</td>
</tr>
<tr>
<td>Sessions (Student Sections pair)</td>
<td>185280</td>
<td>242902</td>
<td>51697</td>
</tr>
<tr>
<td>Max highlights per student</td>
<td>3038</td>
<td>1079</td>
<td>310</td>
</tr>
<tr>
<td>Mean highlights per section (Standard Deviation)</td>
<td>4.2 (3.6)</td>
<td>0.9 (1.4)</td>
<td>1.2 (0.11)</td>
</tr>
</tbody>
</table>
of highlighting is encoded in a vector representation, and we explore a range of representations which we explain shortly. We use the simplest possible model—a linear regression model with the vector representation as the regressor and quiz score as the regressand. Figure 3.3 shows an excerpt of text from one section, whose aim is to summarize the steps of the scientific method and the nature of scientific reasoning. The words in the text are color coded to indicate whether highlighting that word raises (red) or lowers (blue) the model’s prediction of quiz score. Notice that material pertaining to hypothesis testing is associated with better performance and the sentence pertaining to E. coli bacteria is associated with worse performance. Although the E. coli sentence is substantive, it is not the focus of this section of text. Figure 3.2 shows the highlighting patterns of two students for this section. The model correctly predicts the ranking of the two students.

We built models to predict quiz scores from a representation of the highlighting pattern. Separate models were constructed for each section using only the data from students who highlighted at least some words in the section. For this research, we decided to stick to a simple linear model—ridge regression—and to focus on how the highlighting pattern is represented. The results we report are obtained via 10-fold cross-validation. The $L_2$ penalty term was weighted with a coefficient of 0.01, chosen by brief manual experimentation to produce models only slightly different than straight-up linear regression.

In our earlier modeling work using laboratory data [62], we explored a vector representation in which each element of the vector corresponded to one unit of text—either a word, phrase, or sentence—and the element’s value was either binary or continuous. Binary representations indicate whether any character in that span of text was highlighted. Continuous representations indicate the proportion of characters in the span of text that was highlighted. Here, instead of parsing the text by lexical units, we simply blocked the text by a number of characters, with blocks ranging in size from 100 to 15000 characters. We again considered binary and continuous vector representations. Figure 3.4 shows the variance explained by binary and continuous representations, colored in red and green, respectively. The points indicate means across the sections with the standard error bars indicating uncertainty in the estimate of the mean. The results show a clear trend: as the text-block
Figure 3.3: Sample material from one section of the biology textbook, the focus of which is on the scientific method. The intensity of the red (blue) color indicates a model’s prediction of increased (decreased) accuracy on the quiz when the corresponding word is highlighted. This result comes from the model that used the PCA(10%) representation of highlights, and for this section, this model used only the first principal component.

size increases, models better predict scores. We suspect the reason for this improvement is due to the overfitting of the models. There is a tension between more granularity, which can capture subtle differences in highlighting, and fewer parameters, which can prevent overfitting. We ought to have explored the full span of this continuum, but we stopped at blocks of 15000 characters. Nonetheless, for all block sizes, we find that the highlighting pattern reliably predicts score.

In our laboratory study [62], we explored a phrase-level representation that involved manually segmenting the text by phrases, which roughly corresponded to the text delineated by commas, semicolons, and colons. However, it would have been too significant a manual effort to do this segmentation on a larger scale. Instead, we used the NLTK package [42] to divide the sections...
into sentences and constructed a highlighting representation with one vector element per sentence. Neither the binary nor continuous sentence-level representation achieved good performance, as indicated by the black points in Figure 3.4.

We were concerned about overfitting, considering that the smallest data set had only 31 students and the number of model parameters could be greater than the number of data points. To address this concern, we performed logistic principal components analysis (LPCA) to reduce the dimensionality of the highlighting representation. We formed binary vectors with one element per word in a section. Element \( i \) of the vector for a given student was set to 1 if the student had highlighted the word \( i \) in the section. Feeding these word-level vectors into LPCA, we obtained the LPCA decomposition of the vector space and LPCA representation of the highlights for each student. We constructed models using the top \( k \) components for various \( k \).

To address the overfitting issue, we varied the number of components to be proportional to the size of our data set. With \( S \) being the number of students, we expressed \( k \) as a proportion of \( S, k = S/\alpha \), for \( \alpha \in \{1, 2, 3, \ldots, 20\} \). In Figure 3.4, the blue points labeled \( \text{pca}(S/\alpha) \) indicate that increasing \( \alpha \) leads to better score predictions.

In a final series of simulations, we selected \( k \) not based on the size of our data set but on the word length of the section. With \( W \) being the number of words in a section, we chose a percentage \( \beta \) as the dimensionality of the reduced representation, i.e., \( k = \frac{\beta}{100}W \). The purple points in Figure 3.4 labeled \( \text{pca}(\beta) \) show the benefit of decreasing \( \beta \) to obtain the surprising finding that with \( \beta \leq 30\% \), we see a significant boost in the model’s predictive power over previous models. It is reassuring that the precise choice of \( \beta \) does not seem to matter, suggesting that the result is robust.

### 3.3 Results

We express the accuracy of models in terms of the proportion of variance in quiz scores explained by the highlighting pattern. Any non-zero value indicates some explanatory power. Figure 3.4 shows results from a variety of representations, which we will explain shortly.

In all of the above approaches, we find that decreasing the dimensionality of the highlighting
representation is beneficial. This finding could either be due to overfitting issues, as we have speculated, or to the fact that there is a low-dimensional structure in the highlighting patterns. We suspect it is the former, and plan to conduct further investigations optimizing the number of LPCA components based both on $S$ and $W$. Of course, any results we obtain by the present cross-validation methodology will need to be confirmed by tests using another data set; at this point, we cannot entirely trust that true model performance will be as good as is suggested by the best of our cross-validation scores.

We find that with a suitable representation of a student’s highlighting pattern, we can explain about 13% of the variance in their test performance. While 13% is not on an absolute scale a large
fraction of the variance, one must consider the many factors that play into a student’s learning and retention, including their interaction with course materials outside of the textbook (e.g., in class, homework, etc.), their prior knowledge, conditions in which they are reading the text, and their degree of engagement with the current material and past sections. Given these highly influential factors, it’s remarkable that as much as 13% of variance can be explained by highlighting patterns.

We found that the choice of highlighting representation was critical in determining how useful highlights are to predict quiz performance. While it is still an open question what this positional representation is measuring we speculate that students with similar level of comprehension of the material focus on the same phrases, for instance, for the biology section (Figure 3.3) students who understand the material well focus on sentences such as "this method is the testing of hypotheses by means of repeatable experiments" or "theory is a tested and confirmed explanation for observations or phenomena.” while students who lack understanding will focus on sentences that are about "Escherichia coli bacteria". However a more solid theoretical justification is required and possible future work. Nonetheless, the fact that the PCA(10%) representation was superior across all three courses provides some reason for optimism. The fact that it is a fairly compact encoding of the myriad possible highlighting patterns also offers the promise that we may be able to interpret the relationship between these components and course content.

In past work using data produced by laboratory participants [62], we did not find as significant a signal in the highlighting patterns, but it’s a bit difficult to compare the laboratory study to the present study because the laboratory study predicted answers to specific questions, and here we are predicting overall scores. The laboratory study also used a variant of item-response theory which incorporated latent student abilities and item difficulties; these latent factors could supplant some of the signals in the highlighting patterns.

3.3.1 Replication in other data

During the year 2019, we have collected more data with the same settings (More in-depth in Chapter 4). With the new data set, we have conducted the previous analysis to see if the results
are replicated.

Figure 3.5: Proportion of variance ($\rho^2$) in quiz scores explained by a student’s specific highlighting pattern. The bars show one standard error of the mean. Each point along the abscissa corresponds to building a model with a particular representation of the highlighting pattern, as described in the text.

With the new data set we found that with a suitable representation of a student’s highlighting pattern, we can explain about 7% of the variance in their test performance. While the explanation percentage decreased to almost half, we suspect that this is due to having more data in our analysis and there exists some variability across different subjects. Nonetheless, in all of the approaches, we continue to find that decreasing the dimensionality of the highlighting is beneficial. On top of that, the PCA(10%) representation was superior across all courses show some replication of the previous results. Again, of course, any results we obtain by the present cross-validation methodology will need to be confirmed by tests using another data set; at this point, we cannot entirely trust that true
model performance will be as good as is suggested by the best of our cross-validation scores.
Chapter 4

Analysis using the Semantic highlight representation

In the previous chapter, we used a positional representation of highlights to infer student comprehension and retention. In this chapter, we investigate the semantic content of the textbook and the highlighted material with the same goal of inferring comprehension and retention. We conducted our analysis based on data obtained from Openstax Tutor during two semesters in 2019. Students were enrolled in Biology, Physics, Sociology, and History courses. They read sections of their introductory text as part of required coursework, optionally highlighted the text to flag key material, and then took brief quizzes at the end of each section. Semantic representations of text sentences and quiz questions are obtained using BERT embeddings [21]. The similarity of highlighted and non-highlighted text sentences are compared to individual quiz questions to obtain semantic match scores that are used in a regression model to predict quiz question accuracy. As we did in the previous chapter with a positional representation, in this work we use the semantic representation to augment a Rasch model which expresses accuracy in terms of student ability and question difficulty. A key focus of our research here is to determine whether the semantic representation boosts model performance above and beyond the performance obtained by the basic Rasch model. To skip to the punch line, we do observe that semantic representation increases predictive power, both over the basic Rasch model and the Rasch model with positional features. Further, we find some evidence that the semantic representation allows us to predict accuracy not only of factual questions, but questions that require inference from the text (questions that are higher in the Bloom taxonomy).
4.1 Data

We obtained a new set of data from the Openstax Tutor platform. The data were collected from January 1, 2019, through December 31, 2019, and span two academic semesters. In addition to the three course topics available to us in Chapter 3—College Biology, College Physics, and Introduction to Sociology—the new data set included a fourth subject, American History. The data set includes college students and also high school students taking AP courses. The highlighting facility was available for all four subjects for two semesters, providing us with far more data to analyze than in Chapter 3.

It is essential to emphasize that these data were collected in a real-world setting. We had no control over how Openstax Tutor was administered, and thus, how the data were collected. Individual instructors controlled the length, content, and frequency of the assignments using Openstax Tutor. We did not control how instructors valued the assignments, which might have been mandatory, available for extra credit, or an optional supplement. Students in the same class may have completed different assignments due to students missing an assignment or dropping a course. Moreover, students had the freedom to spend as much time as they wanted for the assignments before the deadline and study any other additional material or review of highlights. We have no meta-information about the students since the process was completely anonymous. Therefore, we are unable to report or utilize the demographic information about the student sample. Nonetheless, we can be reasonably sure that since these data came from numerous real classes and real students of various ages, the demographics are diverse.

As we did in the previous analysis, we grouped the data by section. Each section is analyzed independently, and we report mean results across sections. Because the textbooks were electronic, they were revised during the period in which we obtained data. As a result, some sections have multiple versions. We collapsed these revisions together since typically only a few words changed from one version to the next, and it was easy to align the highlighted fragments.

Table 4.1 presents an overview of the new data set. There are a total of 11,134 students, 897
Table 4.1: Overview of 2019 OpenStax Data

<table>
<thead>
<tr>
<th></th>
<th>Biology</th>
<th>Biology AP</th>
<th>Physics</th>
<th>Physics AP</th>
<th>Sociology</th>
<th>History</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
<td>4526</td>
<td>313</td>
<td>4956</td>
<td>671</td>
<td>658</td>
<td>4</td>
</tr>
<tr>
<td>Sections</td>
<td>255</td>
<td>125</td>
<td>284</td>
<td>144</td>
<td>169</td>
<td>33</td>
</tr>
<tr>
<td>Sessions</td>
<td>342693</td>
<td>21040</td>
<td>384672</td>
<td>39143</td>
<td>42683</td>
<td>89</td>
</tr>
<tr>
<td>Sessions with highlight</td>
<td>11396</td>
<td>5559</td>
<td>7695</td>
<td>990</td>
<td>1377</td>
<td>2</td>
</tr>
</tbody>
</table>

distinct sections, and 830,320 sessions, where a session consists of a particular student reading a particular section. We continued to observe the trend of a relative scarcity of highlights during reading in the new data. Students made one or more highlights in only 27,019 of the 830,320 sessions. The low highlighting rate is consistent throughout all of the course subjects, providing more evidence that highlighting in an electronic text is awkward and unfamiliar to students. We still obtained a considerable amount of highlighted data thanks to the availability of highlighting features for both semesters.

Our analyses are based entirely on the sessions with highlights because—as we showed in Chapter 2—students who highlight tend to perform better and including both highlighted and non-highlighted sections would provide a dimension to prediction that we are not investigating here. Our focus here is on whether the specific pattern of highlights is informative as to the student’s comprehension.

While the previous study [24] only used core practice questions for their analysis, in our case we used both core and spaced questions. This will allow us to have more data to train the model. Since we model for each section, we wanted to guarantee a certain amount of data. Therefore, we filtered out sections with less than 50 ~ 100 student who highlighted (differ for different analysis) and less than 5 questions associated with the section. This left us approximately 30 ~ 80 number of sections for our analysis. Similarly with the previous study [24], even though we have both students free form response and multiple-choice response, we only used the multiple-choice response in this analysis. We have yet to have an objective way to evaluate the free form response and it is practically impossible to make human graders evaluate all of the free response.
4.2 Methodology

4.2.1 Background on BERT

To capture the semantic meaning of the phrases, we used a state-of-the-art model that was developed by researchers at Google AI Language, BERT. We provide a brief history of work on BERT in order to motivate the variant that we use to obtain semantic embeddings and assess semantic similarity.

BERT [21] is a pre-trained transformer network [60] that has produced numerous state-of-the-art results in various NLP tasks. BERT’s key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modeling. This effort is in contrast to previous efforts which were limited in that they are unidirectional. The results indicate that a language model which is bidirectionally trained can have a deeper sense of language context and flow than single-direction language models. BERT has achieved a new state-of-the-art performance on the Semantic Textual Similarity (STS) benchmark [13]. RoBERTa [40] showed, that this performance can be further improved by adapting the pre-training process such as training the model longer, with bigger batches over more data; removing the next sentence prediction objective; training on longer sequences; and dynamically changing the masking pattern applied to the training data.

Research has further improved BERT’s performance and efficiency. Sentence-BERT (SBERT) is a modification of the pre-trained BERT network that uses Siamese and Triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity [57]. SBERT adds a pooling operation to the output of BERT / RoBERTa to derive a fixed sentence embedding, where pooling strategies range from using the output of the CLS-token, computing the mean of all output vectors, and computing a max-over-time of the output vectors [57]. The results of SBERT have shown to be superior over the state-of-the-art sentence embedding methods in terms of common NLP task benchmarks. Also, SBERT claims to be computationally efficient compared to BERT in sentence similarity tasks. This is because the BERT requires both sentences to be
On a global scale, many researchers are committed to finding ways to protect the planet, solve environmental issues, and reduce the effects of climate change. All of these diverse endeavors are related to different facets of the discipline of biology. Escherichia coli (E. coli), bacteria, in this scanning electron micrograph, are normal residents of our digestive tracts that aid in absorbing vitamin K and other nutrients. However, virulent strains are sometimes responsible for disease outbreaks. (credit: Eric E. Erbe, digital colorization by Christopher Pooley, both of USDA, ARS, EMU)

The Process of Science Biology is a science, but what exactly is science? What does the study of biology share with other scientific disciplines? We can define science (from the Latin scientia, meaning “knowledge”) as knowledge that covers general truths or the operation of general laws, especially when acquired and tested by the scientific method. It becomes clear from this definition that applying scientific method plays a major role in science. The scientific method is a method of research aimed at steps that include experiments and careful observation. We will examine scientific method steps in detail later, but one of the most important aspects of this method is the testing of hypotheses by means of repeatable experiments. A hypothesis is a tentative explanation for an event, which one can test. Although using the scientific method is inherent to science, it is inadequate in determining what science is. This is because it is relatively easy to apply the scientific method to disciplines such as physics and chemistry, but when it comes to disciplines like anthropology, psychology, and sociology, the scientific method becomes less applicable as repeatable experiments become more difficult.

These areas of study are still sciences, however. Consider anthropology—even though one cannot perform repeatable experiments, hypotheses may still be supported. For instance, an anthropologist can hypothesize that an ancient culture existed based on finding a piece of pottery. He or she could make further hypotheses about various characteristics of this culture, which could be correct or false through continued support or contradictions from other findings. A hypothesis may become a verified theory. A theory is a tested and confirmed explanation for observations or phenomena.

Q: Although the scientific method is used by most of the sciences, it can also be applied to everyday situations. A situation is given below. Using the scientific method try to arrange the given steps in the correct order

We therefore utilize SBERT to obtain sentence and question embeddings which can be compared to obtain semantic similarity scores. Figure 4.1 diagrams how a particular highlighted sentence and a particular quiz question can be compared using SBERT.

We illustrate how effective this framework can be to find text sentences similar to the quiz question. Figure 4.2 is an example extracted from the section “The Science of Biology.” The query question is presented to the student along with multiple-choice alternatives. We encode only the question asked, not potential answers for multiple choice questions. Below the query in the Figure are the five sentences deemed by SBERT to be most related to the question. In this example, the...
Query: What is the name for the formal process through which scientific research is checked for originality, significance, and quality before being accepted into scientific literature?

Answer: peer review

<Top 5 most similar sentences in the section corpus>

1. The process of peer review helps to ensure that the research in a scientific paper or grant proposal is original, significant, logical, and thorough. (Cosine Score: 0.8891)
2. Reporting Scientific Work Whether scientific research is basic science or applied science, scientists must share their findings in order for other researchers to expand and build upon their discoveries. (Cosine Score: 0.8415)
3. Peer-reviewed manuscripts are scientific papers that a scientist’s colleagues or peers review. (Cosine Score: 0.8550)
4. These colleagues are qualified individuals, often experts in the same research area, who judge whether or not the scientist’s work is suitable for publication. (Cosine Score: 0.8253)
5. The introduction refers to the published scientific work of others and therefore requires citations following the style of the journal. (Cosine Score: 0.7918)

Figure 4.2: Question is from the section “The Science of Biology” and the following five sentence are the semantically most similar sentences related to the question

question is about the definition of peer review. When we see the top five related sentences found by this framework, we observe that the most related sentence is basically just a rephrase of the definition. As we go down the list the sentences are somewhat related to peer review or have the word in its sentence. Because highlighting a sentence indicates the student choose to pay more attention to the specific content, we assume that the information whether a student highlighted a sentence high on the list will be useful in predicting the chance of the student getting the question correct.

4.2.2 Representing the semantic similarity between highlights and quiz questions

In this section, we address several methodological decisions we needed to make to fully specify a predictive model. First, we have to decide how to partition the textbook into segments: should we parse by words, phrases, sentences, paragraphs? We must also determine of much of the given segment must be highlighted by the student in order for us to consider the segment to be highlighted. We denote the highlighting or non-highlighting of segment $s$ by the student as $H_s$. We can consider
$H_s$ to be binary or continuous:

$$H^{\text{bin}}_s = \begin{cases} 1 & \text{if any word of the segment is highlighted} \\ 0 & \text{otherwise} \end{cases}$$

(4.1)

$$H^{\text{contin}}_s = \frac{\text{words in the segment that are highlighted}}{\text{number of words in the segment}}$$

(4.2)

where $s \in S$, $S$ is the complete set of segments in a section with $|S| = N$ segments.

For each segment $s$ in a section, we obtain a highlighting score, $H_s$, and a BERT match score to the question $q$, $B(s, q)$.

Students may highlight multiple segments, and these multiple highlights produce multiple BERT scores, which we combine. Next we describe the options for combination. Should we consider on the one sentence with the maximum match score? Or should we consider the average match score? Both the average and maximum contain valuable information that could tell much about the student. The maximum will tell us what the most relevant segment the student attended to whereas the average will tell us the average relevance of segments. One could construct plausible stories about why both these statistics could be relevant for prediction. For example, a student who highlighted randomly or abundantly, the best-matching score may be high but the mean score is expected to be low.

We take advantage of a family of functions that span a continuum from computing the mean to computing the maximum of a set of values. Notice that $\frac{1}{n}||x||_1$ is the same as the arithmetic mean when $x \in \mathbb{R}^n$. Also, $||x||_\infty$ is the same as the maximum element in vector $x$. Our aim is to define a family that ranges from the mean to the max. Our approach is based on a relationship among $\mathcal{L}_p$-norms: $||x||_r \leq n^{\frac{1}{r} - \frac{1}{p}} ||x||_p$. If we apply this inequality with $p = r + 1$ for all $r = 1, 2, \ldots$, then we get the following relation:

$$n^0||x||_1 \leq n^{\frac{1}{2}}||x||_2 \leq n^{\frac{1}{3}}||x||_3 \leq \ldots \leq n||x||_\infty$$

Since we know that $n > 0$ we can also infer the following relation

$$n^{-1}||x||_1 \leq n^{-\frac{1}{2}}||x||_2 \leq n^{-\frac{1}{3}}||x||_3 \leq \ldots \leq ||x||_\infty$$
This sequence of inequalities ranges from the mean to the max with the continuum lying in between. We choose a subset of $p$-norms because—as we explain shortly—not only do the mean and the max provide different information about the student’s highlighting patterns, but so do the $p$ values in between.

For any $p$, we will obtain the following definition of a $p$-match:

$$\text{Match}_{p,q} = \left[ \frac{1}{N} \sum_s B(s,q)^p \right]^{1/p}$$

Now we have to determine how $B(s,q)$ and $H_s$ are combined to predict the performance on question $q$. One way is replacing $N$ for $H_s$ as the following

$$\text{HighlightedMatch}_{p,q} = \left[ \frac{\sum_{k \in S_q} H_k B(k,q)^p}{\sum_{k \in S_q} H_k} \right]^{1/p}$$

where $S_q$ is the set of text segments contained in the section to which question $q$ pertains, and $k$ is an index over text segments. For $H_s^{\text{bin}}$, this equation computes the $L_p$ mean match over the highlighted sections; for $H_s^{\text{contin}}$, this equation computes a weighted combination rule.

We also hypothesize that non-highlighted segments can provide some information of the prediction. For example, the von Restorff effect suggests that highlighting might adversely affect the retention of the non-highlighted material [39]. Thus, we consider the segments that are not highlighted via a second match score for the non-highlighted sections:

$$\text{NonHighlightedMatch}_{p,q} = \left[ \frac{\sum_s (1 - H_s) B(s,q)^p}{\sum_s 1 - H_s} \right]^{1/p}$$

Combining the match scores for highlighted and non-highlighted segments over various $p$-norm scores, we obtain a parameterized linear model for the overall match:

$$\text{OverallMatch}_q = \sum_j \alpha_{q,j} \text{HighlightedMatch}_{p_j,q} + \sum_j \beta_{q,j} \text{NonHighlightedMatch}_{p_j,q}, \quad (4.3)$$

where $j$ is an index over a set of norm values including the mean ($p = 1$) and max ($p = \infty$), and $\alpha_{q,j}$ and $\beta_{q,j}$ are free parameters fit to data.

We have yet to specify (1) how we select a set of $p$ norms, how we represent the highlights (as binary or continuous), and how to segment the text (into sentences or sub-sentence fragments).
4.2.3 Development of the augmented Rasch model

While it is possible to directly calculate the correlation between equation 4.3 with correctness of each question, we were concerned about the nature of information that highlights provide. Models based only on the highlights might succeed not because they determine whether the highlight itself aids the students’ understanding, but because the highlights provide some general information about how skilled or motivated a particular student is. To address this hypothesis, our present work follows a long tradition in the educational data mining community, we use a feature-based regression model. Feature-based regression models consist of two approaches: Performance Factor Analysis (PFA) [51], which hand-crafts features, and Deep Knowledge Tracing, which encodes past history of student performance [54]. In addition to models that incorporate observable features, models such as PFA include latent features—features inferred from but not explicit in the data. The classic latent-feature model in student modeling, item-response theory (IRT), will be used as our baseline. In particular, we use the most basic formulation of IRT, the Rasch model. We augment the Rasch model with additional information that we believe to be related to the dependent variable being predicted. In this work, we augment with features that indicate how a student highlights the text. Highlighting patterns constitute a novel source of side information [62]. This framework combines the latent and observable features into a single interpretable model, which will allow us to evaluate whether highlights offer an orthogonal source of information to student ability and question difficulty.

To formalize, let \( n_s \) denote the number of students given a section with \( n_q \) questions, with \( y_{sq} = 1 \) if the response of student \( s \) to question \( q \) is correct. The Rasch model (also known as the one-parameter logistic) variant of IRT makes the prediction:

\[
P(y_{q,s} = 1) = \text{logistic}(\theta_s - \gamma_q)
\]

where \( \theta_s \) denotes the latent ability of student \( s \), \( \gamma_q \) denotes the latent difficulty of questions \( q \). Since this model does not incorporate any information about the student’s highlighting, we refer this model as the baseline.
Now we construct another model that incorporates information about highlights. We augment the baseline model by incorporating a semantic highlighting representation, OverallMatch\textsubscript{s,q} (Equation 4.3), yielding
\[
P(y_{q,s} = 1) = \text{logistic}(\theta_s - \gamma_i + \text{OverallMatch}_{s,q}).\]  

(4.4)

We perform hierarchical Bayesian inference by placing priors on the parameter vectors $\theta$, $\gamma$, $\alpha$, $\beta$ and estimating hyper parameters of the prior distributions. We specify priors $\theta_s \sim N(\mu_\theta, \sigma^2_\theta)$, $\gamma_i \sim N(\mu_\gamma, \sigma^2_\gamma)$, $\alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha)$, $\beta_j \sim N(\mu_\beta, \sigma^2_\beta)$. All of the feature-based regression models were fit using STAN [11]. We sample four Markov chain Monte Carlo (MCMC) chains each having 4000 samples, and from each chain we remove the first half of samples as burn in. The remaining samples are then averaged together across the four chains to obtain the estimated prediction.

We used two performance measures to evaluate models: area under the ROC curve (AUC) and the area under the precision-recall curve (PRC). AUC represents the degree or measure of separability, indicating how well the model discriminates between classes. Concerns have been raised about using the AUC, claiming that this measure could be misleading when applied to imbalanced classification scenarios [59]. Considering that students have a predominance of correct responses, we considered that imbalance might be an issue, hence decided to use PRC. PRC curves address the imbalance issues by providing an accurate prediction of future classification performance due to the fact that they evaluate the fraction of true positives among positive predictions. PRC plots, on the other hand, can provide an accurate prediction of future classification performance due to the fact that they evaluate the fraction of true positives among positive predictions [59].

To clarify this using an example, consider a case where we have 1000 balanced data for each class in a binary classification and we have a model result with 700 True Positive, 300 False Negative, 160 False Positive, and 840 True Negative. Consider another case where we have an imbalanced data 1000 positive, 10000 negative where the model result with 700 True Positive, 300 False Negative, 1600 False Positive, and 8400 True Negative. Since, the receiver operating characteristic curve (ROC curve) uses the True Positive Rate (\(\text{=True Positive} / (\text{True Positive + False Negative})\)) and
False Positive Rate (=False Positive / (False Positive + True Negative)) it will return the same AUC for both case. However, PRC captures the difference with the Precision term (True Positive / (True Positive + False Positive)), as in the second case this value will be significantly low. In other words, AUC is unaffected by imbalanced data. It normalizes itself with respect to the number of positives and number of negatives. PRC as it shows is sensitive to that.

4.3 Results

We model each section independently and we include only students who make one or more highlights in a section. Our models predict the probability of a correct response to a specific question by a specific student, leveraging the latent ability of the student, the latent question difficulty, and a semantic representation of the highlighted segment.

The semantic similarity scores are obtained by first parsing the entire section into sentences and obtaining the BERT similarity score between each sentence and each question in that section. Since this similarity score would be in the range of \([-1, 1]\), for convenience and interpretability of model parameters, we rescale this score to the range \([0, 2]\), where a larger value indicates greater similarity. We also obtain an \(H\) vector indicating for each sentence whether that sentence had highlighting. From the BERT scores and \(H\) we can compute \(\text{HighlightedMatch}_{p,q}\) and \(\text{NonHighlightedMatch}_{p,q}\) for each question \(q\) and a given norm-\(p\) value.

A critical decision in model building is determining the set of \(p\) values used for the norms. We used \(p\) values of 1, 5, 10, and 100 for both \(\text{HighlightedMatch}\) and \(\text{NonHighlightedMatch}\). We selected these values to optimize the model’s explanatory power. The value 1 is a simple mean of match scores; the value 100 is essentially the maximum of the match scores. Given that the average number of sentences in each section was 104, we ran a quick simulation in which we drew 104 random ‘match scores’ from a uniform distribution \(U(0, 2)\) and computed the \(p\)-norm for \(p\) ranging from 1 to 100 (Figure 4.3). We we plot out a simulation showing the difference in norm values (Figure 4.3) and picked \(p\) values that would be differentiable enough. We observe that there is quite a gap between 1 and 10 but the measured similarity does not change much as \(p\) increases.
4.3.1 Results based on different hold-out settings

We conduct three types of cross-validation analysis. The first type is “hold-out participant and question” which means we randomly select 20% of student-question pairs and regard them as the validation set. In this analysis, the model is trained on all students and all questions, but it wasn’t trained on the particular students answering the particular questions that we have held out. The overall results are shown in Figure 4.4. We observe that the modified IRT model (ability + difficulty + highlight feature) outperforms the baseline; AUC rises from 0.687 (SE 0.004) to 0.719 (SE 0.002) and PRC rises from 0.861 (SE 0.003) to 0.880 (SE 0.002). We also observe that the latent ability provides the least amount of information [AUC: 0.567 (SE 0.009) PRC: 0.797 (SE 0.005)], but this is expected since usually for each section there are many more students than there are questions, so there is less opportunity for the ability parameter to generalize. One surprising
Figure 4.4: Each bar indicates the average of these metrics across all sections and error bars reflect $\pm 1$ SEM, corrected to remove variance due to the random factor [44]. Each point along the abscissa corresponds to combination of different features. The lighter color bars indicate performance when the features are used alone, the darker color bars indicate when the highlight features are added.

The observation is that the highlight feature [AUC: 0.6892 (SE 0.004) PRC: 0.866 (SE 0.002)] does as much as difficulty features [AUC: 0.691 (SE 0.002) PRC: 0.858 (SE 0.003)]. A possible explanation for this is that the highlight features are representing similar information as the question difficulty. However, since the difficulty + highlight combination outperforms both the difficulty and highlight features alone, it is reasonable to suppose that the highlight feature does provide some additional information distinguishing itself to question difficulty. The PRC shows qualitatively similar results.

We next ran an analysis holding out participants, which means we hold on to 20% of the participants are removed from the training set and used to evaluate the model. Results of this analysis are shown in Figure 4.5. In this case, the student ability is not useful because the ability parameter needs to be constrained by data; thus, the AUC based on student-ability alone is near 0.5, an indication that the model cannot predict question accuracy. And the models with ability and other features perform identically to models without ability. Similar to the results with held out participant-questions, the difficulty + highlight combination [AUC: 0.719 (SE 0.002) PRC: 0.876 (SE 0.001)] outperforms difficulty alone [AUC: 0.693 (SE 0.003) PRC: 0.857 (SE 0.002)]. It is thus reasonable to conclude that the highlighting features provide some additional information that can
Figure 4.5: Each bar indicates the average of these metrics across all sections and error bars reflect $+/- 1$ SEM, corrected to remove variance due to the random factor \cite{44}. Each point along the abscissa corresponds to combination of different features. The lighter color bars indicate performance when the features are used alone, the darker color bars indicate when the highlight features are added.

Figure 4.6: Each bar indicates the average of these metrics across all sections and error bars reflect $+/- 1$ SEM, corrected to remove variance due to the random factor \cite{44}. Each point along the abscissa corresponds to combination of different features. The lighter color bars indicate performance when the features are used alone, the darker color bars indicate when the highlight features are added.

be distinguished from question difficulty.

Our final cross-validation analysis involves held-out questions, which means we hold on to 20% of the questions from each section to evaluate the model and train with the rest. Results from this analysis are shown in Figure 4.6. Because the difficulty of questions is determined by
Figure 4.7: The number between the parenthesis indicate the Bloom’s Taxonomy level of questions that were predicted. The lighter color bars indicate performance of the Rasch Model, the darker color bars indicate when the highlight features are added to the IRT model. Error bars reflect +/- 1 SEM, corrected to remove variance due to the random factor [44].

The availability of a question in the training set, the model cannot utilize question difficulty, as reflected by an AUC of 0.5 for the model with only difficulty features, and by the equivalence of models that include and exclude difficulty features. Student ability alone achieves some measure of discriminability [AUC: 0.571 (SE 0.005) PRC: 0.794 (SE 0.002)]; however, adding the highlight feature to ability makes little difference. This finding is consistent with the laboratory study of Winchell et al. [62] where it was found that highlighting features did not boost model performance relative to the baseline model (and in fact did somewhat worse due to overfitting). The highlighting features have value only when the model has no useful information about either ability or difficulty (the blue bars). In this case, the highlighting features do allow prediction slightly above chance.

4.3.2 Model prediction accuracy based on Bloom taxonomy

In this section, we investigate whether the model can predict performance for questions at the five levels of the Bloom Taxonomy; from most concrete to most abstract, these levels are:

1. recall
2. understand
(3) apply

(4) synthesize

(5) evaluate

(6) create

It has been hypothesized in the past that highlighting might be useful for predicting questions at the lowest levels, reflecting simple memorization, but not questions at the higher levels, which require a deep conceptual understanding of the material. We investigated which levels of the Bloom taxonomy benefit from our encoding of highlighting. The results we report are obtained via five-fold cross-validation. As in previous analyses, only students who make one or more highlights in a section are considered for that section. We filtered out sections with less than 50 students who highlighted and less than 5 questions of the level we intend to predict associated with the section. Due to the limited number of questions, we grouped the question levels into three bins instead of the original Bloom’s Taxonomy distinction. The first group contains only the recall questions, the second group contains to understand and apply questions, and the last group contains synthesize, evaluate, and create questions.

Figure 4.7 shows the results for the analysis. Adding highlighting features to the Rasch model improves both AUC and PRC across all levels of the Bloom taxonomy. For recall questions, AUC rises from 0.695 (SE 0.003) to 0.709 (SE 0.003) and PRC rises from 0.879 (SE 0.001) to 0.888 (SE 0.001). For understand and apply questions, AUC rises from 0.678 (SE 0.002) to 0.721 (SE 0.002) and PRC rises from 0.841 (SE 0.001) to 0.869 (SE 0.001). For synthesize, evaluate, and create questions, AUC rises from 0.661 (SE 0.003) to 0.683 (SE 0.003) and PRC rises from 0.821 (SE 0.001) to 0.831 (SE 0.001). The overall predictive power of the models tend to drop for higher levels of the Bloom taxonomy, which we were expecting considering that at higher levels, the complexity of the questions implies that many more factors can come into play in determining student correctness.
4.3.3 Model prediction accuracy with syntactic versus semantic highlighting representations

In Chapter 3, we found that positional features (the pattern of which sentences are highlighted in a given section) are useful for predicting overall quiz performance. In this Chapter so far, we found that semantic features (based on the meanings of the sentences) are useful for predicting correctness of response to specific quiz questions. The two lines of research are based on different data sets. In this section, we attempt to unify the two lines of research by investigating whether the positional features identified in Chapter 3 are useful for predicting correctness of specific questions in our new, larger data set. And if these positional features are useful, are they as useful or more useful than semantic features?

In Chapter 3, the positional representation we found to predict best is to perform principal components analysis on the highlighting pattern and select the 10% of components with the largest eigenvalues. For most of the sections, the top 10% corresponds only to the first component, although for larger sections (with more sentences), we might select two or three components. However, to simplify the present investigation, we selected only the first principal component for each section.
to use for the highlighting representation. The results we report are obtained via five-fold cross-validation. Only students that highlighted one or more letters were included. We filtered out sections with less than 100 students who highlighted and less than 5 questions of the level we intend to predict associated with the section.

Figure 4.8 shows the results for the analysis. For the original Rasch model, the mean AUC is 0.687 (SE 0.005) and the mean PRC is 0.861 (SE 0.003). For the Syntactic IRT model, the mean AUC is 0.700 (SE 0.008) and the mean PRC is 0.867 (SE 0.003). For the Semantic IRT model, the mean AUC is 0.719 (SE 0.005) and the mean PRC is 0.880 (SE 0.003). Based on our results, we conclude that both highlight representations improve prediction performance compared to the baseline (Rasch) model. We also conclude that the model using the semantic highlight representation outperforms the positional highlight representation on both AUC and PRC metrics.

An obvious next step, but sadly one for which we did not have time, is to combine the positional and semantic features into one model. Because these features represent complementary information, we hope to obtain further model improvements with the combination.
Chapter 5

Conclusion and Further Research

To briefly summarize our study, we examined the relationship between student highlighting behaviors and learning performance by focusing on finding the effective highlighting representation. We took two approaches, first we attempt to use highlighting patterns and found that parsing the text into words and reducing the dimension using principal component analysis is most effective in explaining quiz performance. Our second approach is a semantic approach by using BERT comparing highlighted segments to the question and see whether students who highlight segments closely related to the question have a higher chance of getting the question correct. We found that it indeed improves prediction performance when our new semantic representation is added as another variable to the item response theory model. This result was constant across all levels of questions.

Our research is important and novel in four particular respects. First, our results extend across a large sample of students, course topics, and specific content. Second, we move outside a laboratory setting (e.g., [22, 32, 50]) and observe students in an authentic learning environment. Third, we move beyond overall analyses of whether students who highlight score better on quizzes (e.g., [24]) to understand how specific patterns or content of highlights predict comprehension and retention. Finally, our research showed hope that we could expand toward understanding student’s higher-level concepts with better highlight representation. Considering Natural Language Processing research is still actively being developed in the future it might be possible to design better highlight representations.
Our research has several potential limitations and thus several path we could investigate. First, due to the fact that Openstax Tutor selects questions aimed to be at an appropriate level for students, there is some possibility of a confound that yields an optimistic estimate of the utility of highlights. For instance, it’s possible that more motivated students tend both to highlight and to attain a certain level of performance that drives the specific questions being selected. Instead of considering all of the questions identically it might be more accurate to weight the questions based on their difficulty. The most simplistic method is using the bloom level of each question and consider questions in the higher lever as more difficult questions. However, we could be more creative, we could train a IRT model with our data, extract the IRT parameters to obtain the difficulty for each question [34].

Second, we have not used all the potential information in the highlighting patterns: in principle, we could leverage dynamic information about the order in which highlights are made, the time lags between highlights (which indicate the pace of reading), and the deletion of highlights (which presently do not register in our analyses), and compute some measure of informativeness or newness of the sentences that are highlighted. To give a concrete example, we could think about using students reading time as a baseline model. We express reading time by using the average reading time per sentence. This new variable might be related to highlighting since students who highlight would tend to spend more time reading, thus it would be critical to know what different information this new variable provides.

Third, we do not consider individual differences among students except insofar as their highlighting pattern is concerned. Because students will use Openstax resources over the duration of a course semester, we have the opportunity to make multiple observations from the same student and to assemble a profile of that student which ought to provide additional information for interpreting their textbook annotations.

Fourth, for the semantic approach we should more systematically explore several methodological decisions, that we made; in our work with the positional representation, these decisions matter. The assumptions we might question include: whether the correct decomposition of highlights is as
the level of complete sentences and not smaller or larger segments; whether a segment of text should be considered highlighted if any portion is highlighted, as opposed to explicitly representing the fraction of the segment highlighted; whether the summary statistics (i.e., values of p) we selected best capture the distribution of highlighted and non-highlighted match scores. One path we are planning to investigate is considering methods for discourse chunking that are more informed and produce coherent segments such as, TextTiling [30]. TextTiling is a technique for subdividing texts into multi-paragraph units that represent passages, or subtopics. This might be more useful than naively chunking the text into random number of letters.

Finally, in this research, we’ve focused on using highlights to model student comprehension, but highlighting is a rich data source for inferring student interests and foci. We might leverage this fact by, for example, clustering students into interest groups based on similarity of patterns of highlighting, or even group students who show disparate highlighting patterns in order to provoke discussions of what material is important. There is also potential to leverage population highlights as a means of feedback to textbook authors and instructors. If students are highlighting unimportant material or failing to highlight important material from the author’s or instructor’s perspective, perhaps the textbooks should be rewritten or students should be guided to the material that is deemed to be most important.
Bibliography


[59] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3):1–21, 03 2015.


