Inferring Student Comprehension from Highlighting Patterns in Digital Textbooks: An Exploration in an Authentic Learning Platform

David Y.J. Kim[1][0000−0003−4057−0027], Adam Winchell[1], Andrew E. Waters[3], Phillip J. Grimaldi[3], Richard Baraniuk[3], and Michael C. Mozer[1,2]

[1] University of Colorado at Boulder, Boulder, CO, USA
[2] Google Research, Brain Team
[3] Rice University, Houston, TX, USA

Abstract. We investigate whether student comprehension and knowledge retention can be predicted from textbook annotations, specifically the material that students choose to highlight. Using a digital open-access textbook platform, Openstax, students enrolled in Biology, Physics, and Sociology courses read sections of their introductory text as part of required coursework, optionally highlighted the text to flag key material, and then took brief quizzes as the end of each section. We find that when students choose to highlight, the specific pattern of highlights can explain about 13% of the variance in observed quiz scores. 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. Considering the many sources of uncontrolled variability affecting student performance, we are encouraged by the strong signal that highlights provide as to a student's knowledge state.

Keywords: student modeling · textbook annotation · knowledge retention

1 Introduction

Digital textbooks have become increasingly available with the popularity of e-readers and the advent of open-access learning resources such as Openstax. 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 [8]. 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 a compelling example of implicit measures, eye gaze has been used to predict predict mind wandering during reading [4].

In this article, we are interested in highlighting—the yellow marks and underlines that students make in a textbook in order to emphasize material that they perceive as particularly pertinent. Although there is a well-established research literature in educational psychology examining whether highlighting benefits student learning [3, 12], 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. Highlighting is a popular study strategy [5] and students voluntarily highlight because they believe it confers learning benefits [1]. Given that highlights reflect material that students believe to be important, one has reason to hypothesize that highlights could be useful for assessing comprehension.

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. [11] asked participants to read and optionally highlight three sections of an Openstax biology textbook [9], 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. [10] 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 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.

In this article, we aim to integrate the focus of these two previous studies, in order to determine the effectiveness of models that leverage the *pattern of highlights* a student produces to predict quiz performance in an *authentic digital learning environment*. We utilize the Openstax Tutor corpus of Waters et al. and construct a model for each section of text, predicting mean quiz performance for that section based on the entire set of highlights in that section.

2 Data Set and Methodology

Our analyses use data collected from Openstax Tutor [10], an online textbook and learning environment used by students enrolled in authentic advanced high school and college courses. Data were gathered from two semesters during the year 2018 for three subjects: College Biology, College Physics, and Introduction to Sociology. Associated with each course was a textbook. Each textbook is divided into chapters which are further subdivided into *sections*. At the end of each section, students were given the opportunity to answer three *core* questions relating to the section. Questions, which ranged from factual to conceptual, were chosen from a pool of candidates by Openstax Tutor based on the student's ability level. Students could repeat the quiz, each time getting new questions. In addition, students were occasionally asked *spaced-practice* questions that could be associated with any section of any chapter previously studied.

The digital textbook environment provided an annotation facility that allows students to highlight critical material in the text by click-and-dragging the mouse over the material they wished to emphasize. Highlights could also be undone. Highlighting was optional. Some students never used the facility, others used it for only some sections. Figure 1 shows highlighting patterns of two different students for the same section of material.

Our analyses were all conducted by section. For each student and each section, we grouped together all questions answered by the student, both core ques-

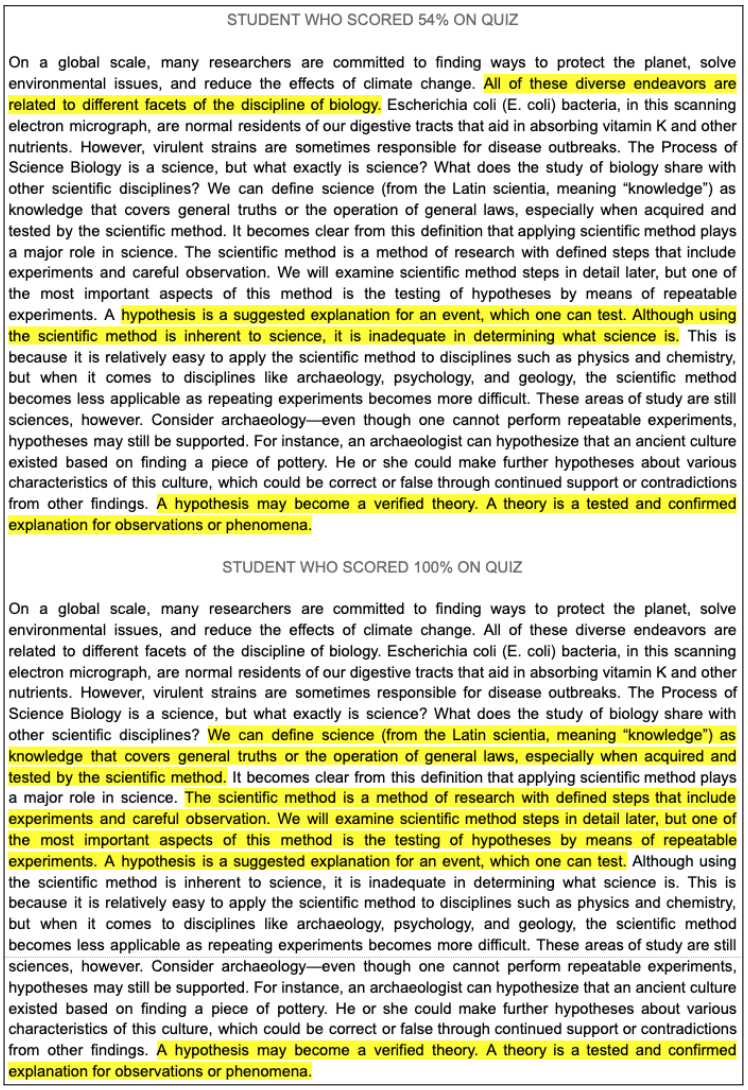


Fig. 1: Samples of student highlighting from one section of the biology textbook, the focus of which is on the scientific method. The student whose highlights are shown in the upper portion of the figure scored 54% on the section quiz; the student whose highlights are below scored 100%.

Table 1: Data set summary

	Biology	Physics	Sociology
Students	1,946	2,421	484
Sections	608	435	114
Sessions (Student-Sections)	185280	242902	51697
Max highlights per student	3038	1079	310
Mean highlights per section	4.2 (3.6)	0.9 (1.4)	1.2 (0.11)

tions 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 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 Results

3.1 Is highlighting associated with higher quiz scores?

We first report on some simple analyses showing that highlighting is associated with higher quiz scores. Although we cannot ascertain a causal relationship, we can eliminate some confounders. Waters et al. [10] conducted a similar analysis via a latent-variable model that included highlighting as a feature. However, they focused on predicting correctness of response to a particular question based on whether or not the corresponding critical sentence in the text had been highlighted. We broaden this investigation to ask where mean accuracy across all questions depends on 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 2a, 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:

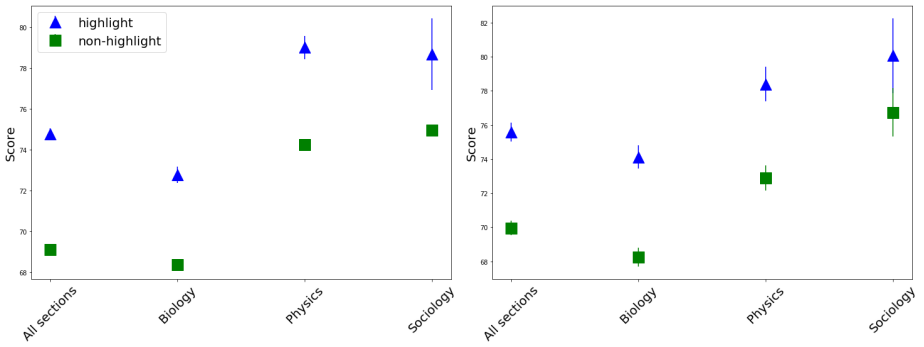
- More diligent students may tend to highlight and more diligent students study hard and therefore perform better on quizzes.
- Whether or not a student highlights may be correlated with the difficulty of the material. For example, a student who is struggling to understand material may not feel confident 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 versus non-highlighted sections. We consider only students who have both highlighted and nonhighlighted sections, and we compute the mean score by student when they highlight and when they do not highlight. This within-student comparison is shown in Figure 2b 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 [6], to infer section difficulty from the student-section observation matrix. We found a miniscule *positive* correlation of 0.02 between difficulty

Table 2: Between student comparison

Topic	# Highlighted Sessions	# Non-highlighted Sessions	t-test
Biology	1,998	32,407	$t(34403) = 9.38, p < 0.01$
Physics	724	40,283	$t(41005) = 6.60, p < 0.01$
Sociology	117	7,405	$t(7520) = 2.06, p = 0.04$



(a) Between student comparison (b) Within student comparison

Fig. 2: Mean scores for highlighted versus non-highlighted sections, by course topic. Error bar indicate one standard error of the mean

Table 3: Within student comparison

Topic	# Students	t-test
Biology	625	$t(624) = 9.30, p < 0.01$
Physics	228	$t(227) = 5.58, p < 0.01$
Sociology	61	$t(60) = 1.61, p = 0.11$

and the probability of highlighting a section which 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 [1].

3.2 Can we predict scores from the specific pattern of highlighting?

The analysis in the previous section 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 highlighting *patterns* help explain scores. 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 of highlighting is encoded in a vector representation, and we explore a range of representations

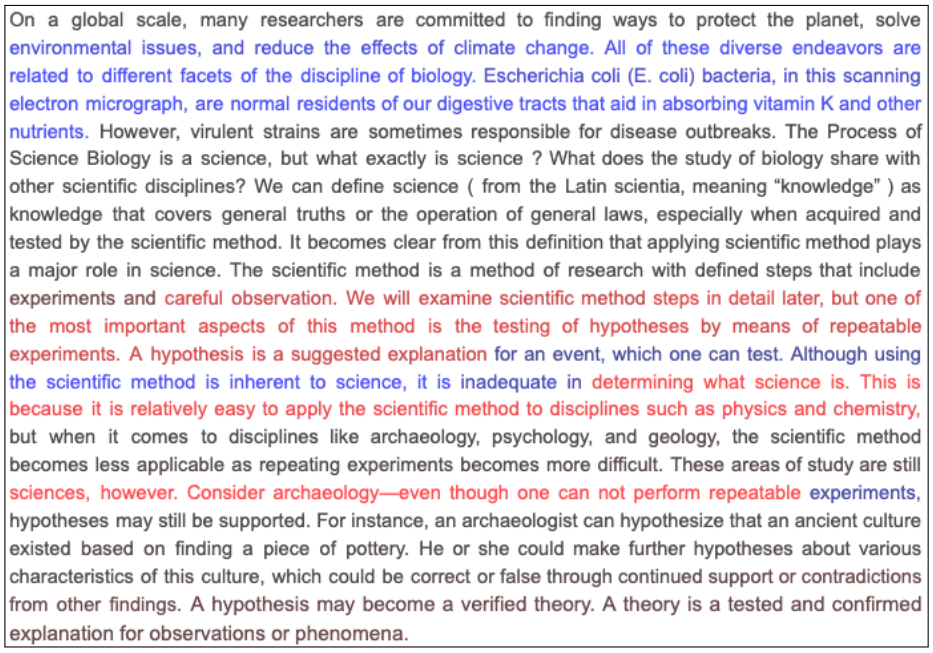


Fig. 3: Sample material from one section of the biology textbook, the focus of which is on the scientific method. The intensity of the red (blue) color indicates a model's prediction of increased (decreased) accuracy on the quiz when the corresponding word is highlighted. This result comes from the model that used the PCA(10%) representation of highlights, and for this section, this model used only the first principal component.

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 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 1 shows the highlighting patterns of two students for this section. The top and bottom patterns are for students who score 54% and 100% on the quiz. The model correctly predicts the ranking of the two students.

We express accuracy of models in terms of the proportion of variance in quiz score explained by the highlighting pattern. Any non-zero value indicates some explanatory power. Figure 4 shows results from a variety of representations, which we will explain shortly. The bottom-line finding is that specific highlight-

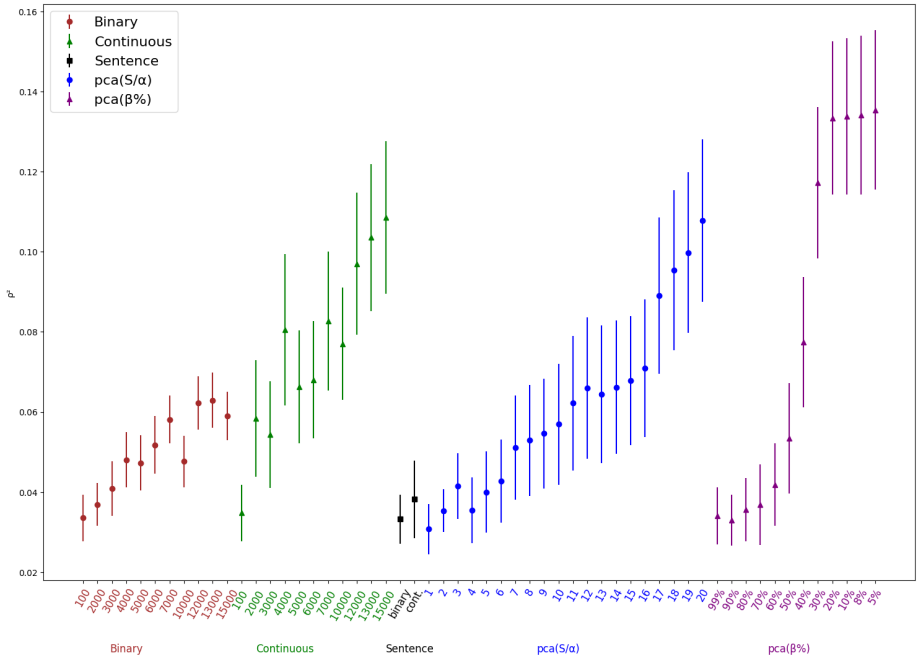


Fig. 4: Proportion of variance (ρ^2) in quiz scores explained by a student's specific highlighting pattern. The bars show one standard error of the mean. Each point along the abscissa corresponds to building a model with a particular representation of the highlighting pattern, as described in the text.

ing patterns can explain about 13% of the variance in scores. This is a fairly impressive effect considering the very large number of factors and influences on a student's performance. For instance, there is some intrinsic variability due to the fact that questions were sampled randomly, and some of the responses were collected immediately after reading while others were collected after a retention period. There is further variability due to the student's momentary state of engagement, the conditions under which they study, and their prior background with the material.

We built models to predict quiz score 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 [11], 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 were highlighted. Here, instead of parsing the text by lexical units, we simply blocked the text by number of characters, with blocks ranging in size from 100 to 15000 characters. We again considered *binary* and *continuous* vector representations. Figure 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 size increases, models better predict scores. We suspect the reason for this improvement is due to 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 [11], 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. However, we used the NLTK package [2] 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 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 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 4, the blue points labeled *pca(S/α)* indicate that increasing α 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 β as the dimensionality of the reduced representation, i.e., $k = \frac{\beta}{100}W$. The purple points in Figure 4 labeled *pca(β)* show the benefit of decreasing β 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 β does not seem to matter, suggesting that the result is robust.

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 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.

4 Discussion

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 choice of highlighting representation was critical in determining how useful highlights are to predict quiz performance. Without theoretical justification for the representation which yielded the best predictions (the top 10% of principal components), we require additional empirical validation to argue convincingly that this representation will also serve us well for other students and other texts. 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 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 [11], 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 signal in the highlighting patterns.

Our research is important and novel in three 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., [1, 11, 7]) 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., [10]) to understand how specific patterns of highlights predict comprehension and retention.

Our research has several potential limitations of this research. 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. Second, we have not used all the potential information in the highlighting patterns: in principle, we could leverage dynamical 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). 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 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. Future research will address these issues.

In this article, 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.

5 Acknowledgement

This research is supported by NSF awards DRL-1631428 and DRL-1631556. We thank Christian Plagemann and three anonymous reviewers for their helpful feedback on earlier drafts of this manuscript.

References

1. Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J., Willingham, D.T.: Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public Interest **14**(1), 4–58 (2013). https://doi.org/10.1177/1529100612453266, https://doi.org/10.1177/1529100612453266, pMID: 26173288
2. Loper, E., Bird, S.: Nltk: The natural language toolkit. In: In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics (2002)
3. Mathews, C.O.: Comparison of methods of study for immediate and delayed recall. Journal of Educational Psychology **29**(2), 101–106 (1938). https://doi.org/https://doi.org/10.1037/h005518
4. Mills C, Graesser A, R.E.D.S.: Cognitive coupling during readingr. J Exp Psychol Gen **146**(6), 872–883 (2017). https://doi.org/doi: 10.1037/xge0000309
5. Miyatsu, T., Nguyen, K., McDaniel, M.A.: Five popular study strategies: Their pitfalls and optimal implementations. Perspectives on Psychological Science **13**(3), 390–407 (2018). https://doi.org/10.1177/1745691617710510, https://doi.org/10.1177/1745691617710510, pMID: 29716455
6. Rasch, G.: Probablistic models for some intelligence and attainment tests (1980)
7. Rickards, J. P., .A.G.J.: Generative underlining strategies in prose recall. Journal of Educational Psychology **67**(6), 860–865 (1975). https://doi.org/https://doi.org/10.1037/0022-0663.67.6.860
8. Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A.: Cognitive tutor: Applied research in mathematics education. Psychonomic Bulletin & Review **14**, 249–255 (2007)
9. Rye, C., Wise, R., Jurukovski, V., Desaix, J., Avissar, Y.: Biology (2016)
10. Waters, A.E., Grimaldi, P.J., Baraniuk, R.G., Mozer, M.C., Pashler, H.: Highlighting associated with improved recall performance in digital learning environment (Submitted)
11. Winchell, A., Lan, A., Mozer, M.C.: Highlights as an early predictor of student comprehension and interests. Cognitive Science p. accepted for publication (2020)
12. Yue, C.L., S.B.K.N.e.a.: Highlighting and its relation to distributed study and students' metacognitive beliefs. Educ Psychol Rev **27**, 69–78 (2015). https://doi.org/https://doi.org/10.1007/s10648-014-9277-z