# Textbook Annotations as an Early Predictor of Student Learning

Adam Winchell[1], Andrew Lan[2], Michael Mozer[1,3]

[1]University of Colorado Boulder

[2]University of Massachusetts Amherst

[3]Presently at Google Brain

**Abstract**

When engaging with a textbook, students are inclined to highlight key content. Although students believe that highlighting and subsequent review of the highlights will further their educational goals, the psychological literature provides little evidence of benefits. Nonetheless, a student's choice of text for highlighting may serve as a window into their mental state—their level of comprehension, grasp of the key ideas, reading goals, etc. We explore this hypothesis via an experiment in which 400 participants read sections from a college-level biology text, briefly reviewed the text, and then took a quiz on the material. During initial reading, participants were able to highlight words, phrases, and sentences, and these highlights were displayed along with the complete text during the subsequent review. Consistent with past research, the amount of highlighted material is unrelated to quiz performance. However, our main goal is to examine highlighting as a data source for inferring student understanding. We explored multiple representations of the highlighting patterns, and built probabilistic and matrix factorization models to predict quiz performance and highlighting patterns, respectively. We conclude that: highlights are predictive of quiz performance for new students and unanswered questions for existing students, and highlighting patterns can be predicted with relative ease. Our long-term goal is to design digital textbooks that serve not only as conduits of information into the mind of the reader, but also allow us to draw inferences about the reader at a point where interventions may increase the effectiveness of the material.

## 1 Introduction

Educational data mining is premised on the assumption that we can collect sources of data that will provide insight into students' *knowledge state*—the degree to which they understand and can apply specific concepts and facts. Typically, such data are first observed when students practice solving problems or taking quizzes.

There can be a long lag between the first exposure to new material and observations of students' performance. For example, in a traditional classroom where a student is assigned a reading from a printed textbook and then takes a quiz days later in class, an opportunity has been lost to perform early interventions. With the advent of electronic texts, the opportunity exists to collect data from students during their early exposure to unfamiliar material, and if knowledge state can be inferred from data collected during reading, interventions can be performed early. To model engagement and comprehension, an obvious source of information is the gaze pattern of a student reading a textbook [27]. However, reliable gaze data are quite difficult to collect in a naturalistic setting. Fortunately, explicit behavioral measures are often available: students voluntarily highlight sections of text and write notes in the margins [5].

The goal of our work is to explore two hypotheses: (1) that textbook highlights—and more generally annotations such as marginal notes—can be used to predict comprehension, as assessed by a follow-up quiz, and (2) that what a student highlights in one section of text can be predicted from what they've highlighted in other sections. To understand the value of highlights as a data source, we conducted an experiment in which participants read and highlighted sections of an electronic textbook, and then took a delayed quiz. We used the observed highlights to predict quiz scores using feature-based regression models [26] that extend on item-response theory [36]. We used observed highlights from two passages to predict the highlights in a third passage using matrix factorization models from the collaborative filtering literature [40].

To summarize our results, highlights help predict quiz performance when we hold out random participants from model training or when we hold out a random subset of questions from each participant (hereafter, *participant-questions*), but not when we hold out random questions from all participants. However, the benefit of adding highlights to predict quiz performance is modest. In exploring various representations of the highlights, we found that coding them in terms of word primitives, i.e., as a feature vector indicting which words are highlighted, achieves the best predictions. We also found that an individual's highlighting pattern informs predictions of what they would highlight elsewhere. Nonetheless, we found that the value of highlights for predictions is limited.

The rest of the paper is organized as follows. In Section 1.1, we review how highlighting data have been collected and used in electronic learning platforms, and the impact of highlighting on learning. In Section 2, we describe our experiment used to gather the highlighting and performance data. In Section 3, we explain the different ways to encode the highlighting data. In Sections 4 and 5, we describe modeling and results. In Section 6, we conclude with a discussion of the future use of highlighting data.

### 1.1 Text Annotations as a Study Tool

Many modern electronic learning platforms allow highlights to be recorded and aggregated across students; we give a brief overview of these systems and their utilization of highlights. *Open-corpus* systems allow users to browse and annotate any content on the web, whereas *closed-corpus* systems have fixed content common to all students. AnnotatED[12] is an open-corpus system that offers annotation tools and adaptive naviga-

tion support (i.e., directing users to resources related to their queries based on the collective knowledge of a community of users). Another open corpus system, *Hypothes.is*, allows users to create annotation layers for different purposes—personal, classroom, public—with the ability to share and discuss their annotations. Aggregated annotations can be used to provide summaries of text, links to relevant content, and personalized document clustering [9]. OpenStax [38] is closed-corpus system that offers open-access college-level introductory textbooks. We are most interested in closed-corpus systems such as Openstax because when many students highlight the same text, there is an opportunity to perform collaborative filtering, i.e., using the data from the population to make predictions about individuals. At present, many of these platforms allow students to highlight, offering an opportunity to mine the data to improve student understanding and retention.

The effectiveness of highlighting as a study strategy has been examined by psychologists, the results of these studies have been mixed (see [11, 28] for recent reviews of the field)[1]. The literature focuses on two experimental designs: those in which participants actively make highlights and those in which participants passively read highlights. Experiments that studied active highlighting, where the participant received no training on how to highlight, have showed no positive relationship between highlighting and a student's exam performance [13, 16, 17]; in fact, one study [34] found highlighting to be detrimental when students were required to draw inferences from the text. In general, students have little idea how to highlight the material that will achieve improved performance [30, 37, 41].

In a few situations, highlighting can provide benefits. First, text which is pre-highlighted by an informed instructor can guide a student to focus on key content [15, 29]. Second, restricting highlighting to encourage consideration of the material—e.g., permitting the student to highlight only one sentence per paragraph—can support understanding [37]. These two examples illustrate that regardless of the difference between active versus passive highlighting, training students how to highlight effectively can improve the potency of the technique [22].

Based on current evidence, one might recommend to students that they simply stop highlighting. However, given students proclivity to highlight, we can leverage highlights as a data source to draw inferences about individuals' knowledge states. We conducted an experiment to collect a data corpus consisting of about 400 students reading and highlighting the same three passages of text.

## 2 Experiment

Participants read passages from a college-level biology textbook. They later reviewed the passages, and then took a short quiz generated from factual material from the passages. During initial reading, participants were allowed to highlight portions of the text (words, phrases, or sentences). During the review phase, these

[1] Underlining and highlighting are treated as equivalent techniques in the literature, and thus results between the two are shared. See [13] for a study that found no difference between the techniques.

highlights were displayed inline with the text. To encourage highlighting, participants were informed that highlights made during the reading phase would be presented along with the full text during the review phase, and that the review phase would be sufficiently brief that a complete re-reading of the text would not be feasible.

## 2.1 Methodology

### 2.1.1 Participants

Participants aged 18 and above were recruited from Amazon Mechanical Turk. A total of 400 people completed the experiment and were paid $3.60. Data from nine participants was discarded. The experiment took 25-30 minutes to complete. To incentivize attention to the task, participants were told that they would be entered into a raffle for a bonus prize of $15.00, with the number of entries equal to the number of correct responses to the quiz questions.

After testing 198 participants, we became concerned that some minor details of the experiment might be influencing results. Thus, we tested the next 202 participants using a slightly altered version of the experiment. We will refer to these two versions as Condition A and Condition B. Of the nine participants removed from the study, six in Condition A reported that they were unable to use the highlighting functionality in their web browser and three in Condition B indicated they were familiar with the experiment material (despite having indicated no familiarity in advance of the experiment).

### 2.1.2 Materials

Three passages were selected from the Openstax *Biology* textbook [38]. The passages were chosen with the expectation that they 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. The passages were shown in this order for all participants. Twelve factual quiz questions were generated by turning specific sentences from the passages into fill-in-the-blank (hereafter, *FIB*) questions. Three questions are drawn from the first passage, four from the second passage, and five from the final passage. These twelve questions were transformed into twelve additional multiple choice (hereafter, *MC*) questions, each question comprised of the correct response and three lures as alternatives.

In scoring, all questions had equal weighting, and we computed a *normalized quiz score* specifying the probability of a correct answer. For judging FIB response correctness, a liberal criterion was used: a response is considered correct if the edit distance between the actual and correct responses is less than 25% of the length of the correct response.

### 2.1.3 Procedure

The experiment is divided into four phases: *instructions*, *reading*, *review*, and *quiz*.

In the instruction phase, participants were given the structure of the experiment, the makeup of the quiz, and were encouraged to highlight because the highlights would be available during the brief review phase. Participants were required to maintain focus on the experiment window, because Amazon Mechanical Turk workers tend to multitask. In Condition A, participation was terminated if the experiment window defocused twice. Because some applications running on the participants' computer could accidentally defocus the window, in condition B we eliminated the termination constraint and replaced it with a requirement that the experiment window be full screen to minimize distractions. In Condition A, we did not screen participants to inquire about their background in biology, but in Condition B, we asked participants if they had taken a college-level biology course in the previous three years and to not participate in the experiment if they had. To ensure that participants were able to anticipate the nature of the text and questions, in Condition B we added a sample paragraph and question to the instructions.

In the reading phase, participants were presented with the three passages sequentially. In Condition A each passage was on the screen for five minutes, in Condition B each passage was displayed for six minutes; the increase in time between Conditions was to alleviate participant concerns of feeling rushed during the reading of the more technical passage. During the reading phase, participants were allowed to highlight text by selecting one or more words using the mouse (highlighting a portion of a word, highlights the full word). If the selected text exactly corresponds to an existing highlight, the highlight is deleted. If the selected text captures any portion of an existing highlight but extends beyond it, the existing highlight is expanded to include the new selection. A single selection of the text may highlight more than one sentence at a time, but cannot cross paragraph boundaries.

In the review phase, participants were presented with the same three passages sequentially, displayed along with any highlights made during the reading phase, each passage on the screen for one minute. Additional highlights were not allowed during the review phase.

During the reading and review phases, a timer at the top of the screen indicated time remaining for the current passage. After the timer expired, the screen blanked and displayed a message describing the next step of the experiment (either the next passage or the next phase of the experiment). Throughout the course of the experiment, a progress bar was displayed at the bottom of the screen that indicated the current proportion of the experiment completed.

In the quiz phase, participants first answered the 12 FIB questions followed by the 12 MC questions. Questions were randomized within question type, as determined by the Condition. In Condition A, the order was randomized for each participant. In Condition B, questions were blocked by passage, maintaining the order in which the passages were read, but randomized within block.[2] At the end of the quiz phase, we asked participants in Condition B whether in retrospect the material was familiar to them. Three participants

[2]The reason for blocking questions was to better control the time between reading of the passage and the quiz.

Table 1: Distribution of response correctness on multiple choice (MC) and fill-in-the-blank (FIB) versions of a question.

| | MC Incorrect | MC Correct |
|---|---|---|
| FIB Incorrect | 0.302 | 0.352 |
| FIB Correct | 0.0580 | 0.288 |

were eliminated due to their reported familiarity with the material. In Condition B, we also asked several questions relating to the participants' perceived effectiveness of highlighting.

## 2.2 Preliminary Analysis

At the end of the experiment, participants in condition B were asked "Do you consider highlighting an effective study strategy?" The proportion of words in the text highlighted by participants was related to their response as indicated by a one-way ANOVA: those responding no, sometimes, and yes highlighted an average proportion of 0.21 (SE=0.04), 0.31 (SE=0.02), and 0.32 (SE=0.04) words, respectively ($F(2, 198) = 3.24, p = 0.041$), consistent with their beliefs about highlighting effectiveness. More surprisingly, their beliefs were also correlated with quiz score: those responding no, sometimes, and yes attained quiz scores of 0.36 (SE=0.03), 0.44 (SE=0.02), and 0.46 (SE=0.02), respectively ($F(2, 198) = 4.93, p = 0.0081$). This positive relationship between survey responses and quiz score contrasts with a negative relationship observed by Yue et al. [42]. Their experiment was remarkably similar in structure to ours, and differed primarily in that their retention intervals were a week long, they collected paper highlights, and their participants were college undergraduates, not Amazon Mechanical Turk workers. Perhaps the difference in results is due to the populations: whereas college undergraduates have recent experience highlighting, Amazon Mechanical Turk workers may not.

A two-way ANOVA was run to examine the effect of experiment Condition (A versus B) on quiz performance on the different passages. There was a significant interaction between the experiment Condition and overall quiz performance ($F(1, 1165) = 11.47, p < 0.001$). However, there were no interactions with other factors (such as passage number, number of highlights made, etc.). Therefore, in all subsequent analyses, we combined data from the two Conditions.

Moving on, we now examine the relationship between highlighting and quiz performance. Fig. 1 shows a scatter plot of the proportion of sentences highlighted and the normalized quiz score, with each point in the plot corresponding to a single participant. As shown along the top margin, the proportion of sentences highlighted is a unimodal distribution with a mean of 0.38. The normalized quiz score, shown along the right margin, is also unimodal with a mean of 0.49. Although the scatter plot suggests no functional relationship between the amount of text highlighted and quiz performance, the correlation coefficient is 0.17 ($p < 0.001$). This correlation drops to 0.079 ($p = 0.15$) when participants that did not highlight are removed from the

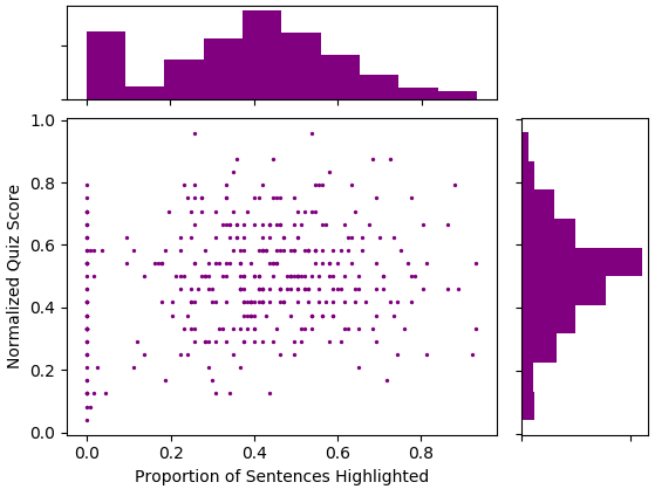


Figure 1: Scatter plot of proportion of sentences highlighted versus normalized quiz score for each participant. The marginal distributions are shown above and to the right of the scatter plot.

analysis. Breaking down the results from the quiz further, Table 1 shows the distribution of response correctness on MC and FIB versions of a question.

Finding no correlation between quantity of text highlighted and quiz performance, we turn to simple models that leverage domain knowledge, specifically, our knowledge of which sentence in the text contains the critical information for a given quiz question. We tested the correlation between highlighting a critical sentence and improved performance on the corresponding quiz question. Analyzing the fill-in-the-blank (FIB) and multiple choice (MC) questions separately, a two-tailed matched-sample $t$-test indicates that MC quiz scores are significantly higher for those who highlighted the critical sentence than for those who did not (0.69 versus 0.57, $t(11) = 6.11$, $p < 0.001$, $d = 0.70$). A marginal effect in the expected direction was also found for FIB by those who highlighted the critical sentence versus those who did not (0.36 versus 0.32, $t(11) = 1.73$, $p = 0.11$, $d = 0.19$).

## 3 Highlight Encodings

Turning from summary statistics of highlighting behavior, we now examine the specific patterns of highlights and their relationship with performance. To set up this examination, we first address the encoding of highlighting data. Fig. 2 presents an example of three participants' highlights of one paragraph of text. As these examples make clear, there is diversity in the manner in which individuals highlight. Highlights demarcate blocks of text ranging from single words to phrases to complete sentences.

In all analyses, we ignore the time course and sequence of text selections and deletions that the participant

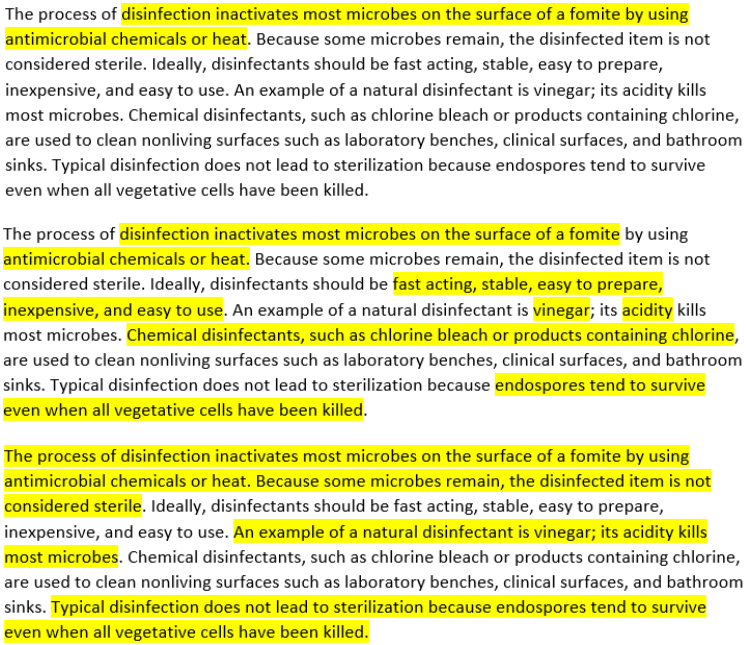


Figure 2: A paragraph of text as highlighted by three randomly selected participants.

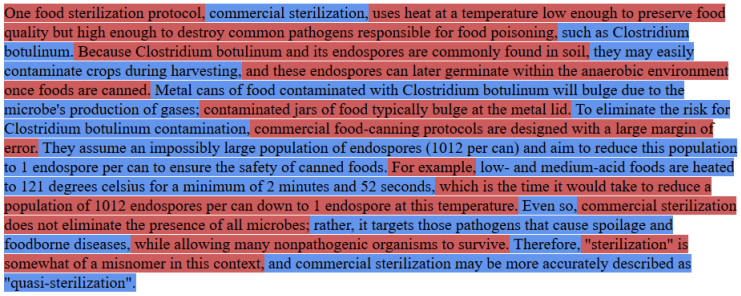


Figure 3: Example of the fragment representation where the alternating colors signify the different fragments.

made, and instead consider only the terminal highlighted state of each passage. The three passages contain 117 sentences delineated by periods, exclamation marks, and question marks. We can parse the highlights at the granularity of a *sentence*, i.e., we will count a sentence as highlighted if any word in the sentence is highlighted. We can also parse the highlights at the level of *fragments*, which are phrases delineated by commas, semicolons, colons, etc. The inclusion or exclusion of commas as segment boundaries was subjective; as an example, our strategy was to exclude commas that were used to delineate lists of items. Fig. 3 gives an example of the fragment partitioning. Our subjective partitioning of phrases yields 235 fragments total; a fragment is counted as highlighted if any word in the fragment is highlighted. Finally, we can also parse highlights at the granularity of individual *words*. The three passages contain 2291 word tokens altogether.

The highlighting pattern of each participant can be described as a binary feature vector whose length depends on the lexical elements. The high-dimensional word-level representation captures the exact pattern of highlights, but using this representation in a regression model introduces many free parameters; the low-dimensional sentence-level representation loses some detail in an individual's highlighting pattern but supports a more compact regression model. Our intermediate level representation, of fragments, was coded manually and based on our parsing intuitions. To provide alternative intermediate-level representations, we examined two automatic methods.

First, we considered an unsupervised method for binary data, logistic principal components analysis (LPCA) [7, 21]. Traditional principal components analysis (PCA) seeks to find a small set of orthogonal components that represents the original data. LPCA primarily differs from PCA in that observations are assumed to be drawn from a Bernoulli distribution rather than a Gaussian. LPCA was applied to the word representation to reduce its dimensionality from 2291 to 120. The value 120 was chosen to be in line with the dimensionality of the sentence representation, and it also preserved 99% of the variance in the data. Using the algorithm variant described in [20], which was well suited due to its interpretability and implicit regularization, we performed 5-fold cross validation to tune their hyperparameter $m$ ($m$ is is used to approximate the natural parameters of the saturated model) for $m \in [1, 10]$, yielding $m = 7$.

Second, we considered several representations based on two traditional natural-language processing methodologies. We used constituency and dependency parse trees to segment the passages into grammatically distinct elements. Neither of these approaches performed well, and we therefore relegate our discussion of these approaches to the Appendix.

# 4 Quiz Prediction

We now turn to the goal of predicting quiz performance given a representation of an individual's highlights. We will use feature-based regression models, following a long tradition in the educational data mining community. Feature-based regression models include performance factor analysis (PFA) [31] and deep knowledge tracing [35]; these two approaches differ in that features are hand-crafted in the former and learned from the

data in the latter. In addition to features that encode past history of student performance, some modeling has incorporated side information, i.e., information not directly related to the dependent measure being predicted. Examples of side information include viewing times and requests for hints [e.g., 8, 43]. Highlighting patterns constitute a novel source of side information. 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 [3], forms the backbone of our regression approach. Item-response theory combines the latent and observable features into a single interpretable model, which will allow us to discern how much leverage highlights give in predicting quiz performance.

To formalize, let $n_P$ denote the number of participants given a test with $n_I$ items, with $y_{pi} = 1$ if the response of participant $p$ to item $i$ is correct. The one-parameter logistic (1PL) variant of item-response theory makes the prediction:

$$Pr(y_{pi} = 1) = \text{logistic}(\alpha_p - \delta_i),$$

where $\alpha_p$ denotes the latent ability of individual $p$, $\delta_i$ denotes the latent difficulty of question $i$.

To this basic model, we incorporated several additional variables. First, because we tested participants in two slightly different conditions of the experiment (explained earlier), we incorporated a binary variable, $e_p$, indicating the experimental condition participant $p$ was assigned to, with values 0 and 1. Second, since we are predicting responses to both MC and FIB variants of a question, we could treat the two sets as independent, however one would expect MC and FIB variants of a question to have correlated accuracy. Still, one would also expect MC variants to be easier. To capture both of these expectations, for each of the $i \in \{1, ..., 24\}$ questions, we separately encoded binary question format (MC versus FIB), $f_i$, and question content (1–12), $c_i$. This encoding results in an additive model for format and content:

$$Pr(y_{pi} = 1) = \text{logistic}(\alpha_p - \delta_{c_i} + \nu_{f_i} + \beta_{e_p}), \tag{1}$$

where $\nu_f$ and $\beta_e$ are free parameters associated with question format $f$ and experimental condition $e$, respectively. We refer to this model as the *baseline*, because it incorporates no information about participant highlighting.

Rather than estimating model parameters $\{\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\nu}, \boldsymbol{\beta}\}$ directly, we perform hierarchical Bayesian inference by placing priors on these parameters and estimating hyperparameters of the prior distributions. We specify priors as: $\alpha_p \sim N(\mu_\alpha, \sigma_\alpha)$, $\delta_c \sim N(\mu_\delta, \sigma_\delta)$, $\nu_f \sim N(0, 2.5)$, and $\beta_e \sim N(0, 2.5)$. To avoid identifiablility issues, $\nu_f$ and $\beta_e$ were used to identify the model (see [2] for advice on dealing with identifying models). All of the feature-based regression models were fit using STAN [6]. We sample four MCMC chains each having 2500 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 a prediction.

Given our baseline model, it is simple to incorporate highlights and associated parameters. We augment the model by incorporating a highlighting representation vector, $\boldsymbol{h}_p$ for participant $p$ and an associated

highlight coefficient vector $\boldsymbol{\omega}_{c_i}$, yielding:

$$Pr(y_{pi} = 1) = \text{logistic}(\alpha_p - \delta_{c_i} + \nu_{f_i} + \beta_{e_p} + \boldsymbol{\omega}_{c_i}\boldsymbol{h}_p^{\mathrm{T}}). \tag{2}$$

The hyperparameters of the baseline model are re-used with priors on the highlighting coefficients being $\boldsymbol{\omega}_{c_i} \sim N(\mu_\omega, \sigma_\omega)$.

## 4.1 Performance Metrics

We evaluate models with two performance measures: Area Under the ROC curve (AUC) and prediction accuracy (PA) [32]. AUC measures how well a model discriminates correct from incorrect quiz answers. AUC was computed by merging all the predictions in the evaluation set. AUC typically lies in the range between 0.5 (no ability to discriminate correct from incorrect) to 1.0 (perfect discrimination). PA is expressed in terms of proportion model predictions that match the student outcome. We use a threshold of 0.5 on the model output probability to distinguish predictions of student correctness and error.

Using these two metrics, we perform cross validation to assess the effectiveness of highlights as a data source for quiz prediction. As with any modeling problem involving participants answering questions, we have multiple options for how to perform the cross validation splits: we can hold out participants, hold out questions, or hold out participant-questions (i.e., hold out a random subset of questions from each participant). Holding out participants or questions allows us to anticipate how our models will fare for new participants and questions, respectively. In this case the associated parameters, $\alpha_p$ and $\delta_{c_i}$ respectively, in Equation 2 are uninformative, the sampled parameter values obtained during training are unchanged during the prediction process. Holding out participant-questions corresponds to the scenario where we have response data from previous participants, and also limited responses from a participant whose later responses we wish to predict [14, 39]. We use 10-fold splits for the held out participants and participant-questions, and 12-fold splits for the held out questions.

## 4.2 Simulation Results

The specific information to answer each quiz question is contained in a critical sentence in the text; for example, the quiz question might ask about a technical term which was defined in the text. As a preliminary analysis, we constructed a version of the model Equation 2 which included only highlights for the critical sentences.

Fig. 4 shows the weight coefficients $\boldsymbol{\omega}_{c_i}$ for the 12 questions on the 12 critical sentences when the model is trained on the complete data set. The main diagonal of this array, which represents the coefficients between the matching sentence and question—indicates that for the most part highlighting the critical sentence increases the model's probability of answering the question correctly, in contrast to the off diagonal coefficients which roughly have a mean of zero. The specific pattern of coefficients is often interpretable in terms of how we constructed the sentences. For example, consider critical sentence 6 and question numbers 6

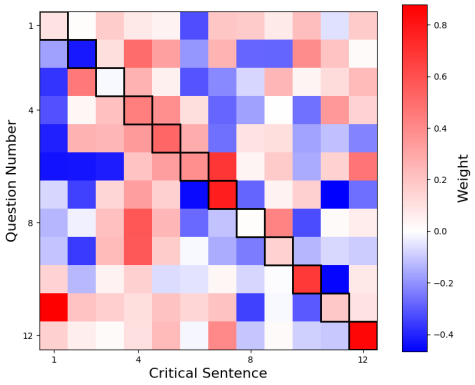


Figure 4: Heatmap of the highlighting weights $\boldsymbol{\omega}_{c_i}$ of Equation 2 using critical sentences. The black squares indicate which critical sentence was used to generate the associated quiz question.

and 7. Question 6 asks about autoclaves which rely on raising temperatures, whereas question 7 asks about lyophilization which relies on freezing. The similarity in the vocabulary—both questions concern themselves with temperature changes—was exploited in the MC versions of these questions, and thus each term was used as a lure for each other. This example illustrates the complexity of analyzing highlights: the relative importance of each highlight is very dependent on the content being tested.

We now move on to the cross validation analyses that utilize the highlight encodings specified in Section 3. Fig. 5 shows a visualization of the average highlighting weights $\boldsymbol{\omega}_{c_i}$ of Equation 2 trained on the sentence representation for the participant-question cross validation. Fig. 5 has one column per sentence in the text, and we have drawn black outline squares to indicate the critical sentences associated with each question, corresponding to the black squares in Fig. 4. Comparing Figs. 4 and 5, it appears that the models trained on the full text still recover the relative importance of the critical sentences in quiz prediction and other similar trends. The coefficients on certain questions are larger than for other questions, as indicated by the horizontal bands in Fig. 5. These bands—as quantified by the mean absolute weight magnitude—are correlated with participant accuracy ($\rho = 0.548, p = 0.065$). From this figure alone, it is impossible to determine whether the highlights are serving a predictive role or whether the highlighting features are being overfit and merely provide a bias on the predicted accuracy.

Fig. 6 shows cross validation performance of models trained on the four different highlight representations (see Section 3). The left graph shows the model's ability to discriminate correct from incorrect quiz responses, quantified by AUC, and the right graph shows prediction accuracy, quantified by proportion correct. The horizontal orange line indicates the performance of the baseline model (Equation 1) which does not utilize highlighting features. Shading represents $\pm 1$ standard error of the mean. (We remove variability arising from the data splits themselves, as suggested by [25], which allows one to compare performance across conditions

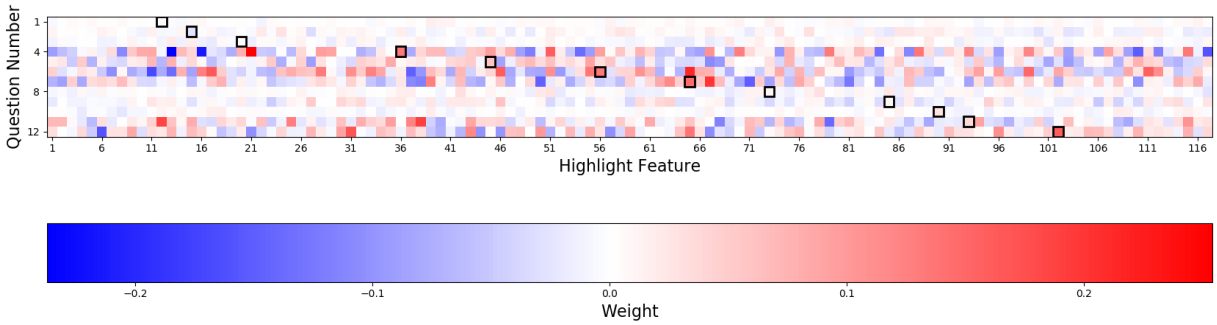


Figure 5: Heatmap of the average highlighting weights $\boldsymbol{\omega}_{c_i}$ of Equation 2 trained on the sentence representation for the participant-question cross validation. The black squares indicate which highlight features were used to generate each quiz question.

by examining overlap of error bars.) The four representations are ordered by granularity, i.e., the number of features in the representation. The finer-grain representations on the right, which indicate word or fragment highlights, outperform the coarser-grain representations on the left, which indicate sentence highlights or a reduced-dimensionality principal-components representation. One would expect this result with sufficient data because the finer-grain representations support more complex models, but we were uncertain a priori whether the number of participants in our study would be sufficient to justify the more complex models.

In Fig. 6, performance is evaluated with held out participant-questions. The models thus have some information about each participant and some information about each question and simply need to fill in the missing cells. We consider this scenario the most realistic in our educational context because in a typical course, we will have data from the students who have taken the course previously, providing information about all questions, and as the course progresses, we will accumulate data from the particular student whose performance we wish to predict. Nonetheless, we can conduct cross validation studies holding out participants and holding out questions.

Prediction on unseen participants (Fig. 7) yields a pattern of results similar to that with unseen participant-questions. There is a clear benefit for the fragment- and word-highlight models, indicating that the pattern of highlights that predicts quiz performance is not participant-specific. This finding is encouraging because it implies that even without prior data on a participant, we can use highlighting to predict memory for text material. Consequently, a participant's comprehension could be assessed online as they first engage with a text.

Prediction on unseen questions (Fig. 8) does not benefit from *any* of the highlighting representations, indicating that the pattern of highlighting attended to by the model is specific to the questions used for training the model. Although successful transfer to new questions would be ideal, instructors often re-use a set of (factual) questions from year to year. One possible approach to obtaining generalization across questions would be to encode questions not by their unique index but by semantic features, allowing for a

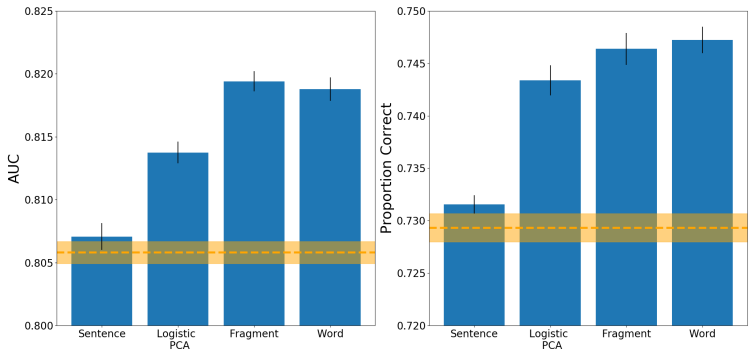


Figure 6: Results for the hold out participant-questions cross validation split using the feature-based regression model for the task of predicting quiz score. The orange line is the baseline result.

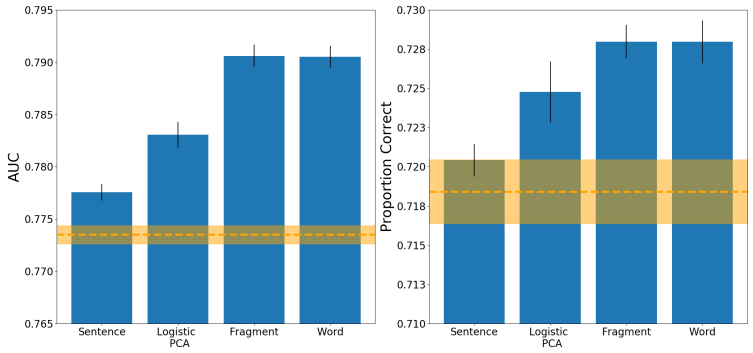


Figure 7: Results for the hold out participants cross validation split using the feature-based regression model for the task of predicting quiz score. The orange line is the baseline result.

natural similarity metric between questions and between a question and the text.

The result on unseen questions suggests that the coefficients on the highlight features do not transfer to questions outside the training set. Consequently, one might suppose models would be even more accurate if the highlight coefficients, $\boldsymbol{\omega}_{c_i}$, were not tied with hyperpriors. (The hyperpriors impose a weak constraint among coefficients for different questions.) Indeed, when we remove the hyperpriors on per-question coefficients, we find a small improvement in model predictions. For example, on the fragment representation of highlights and held out participants, AUC rises from 0.791 (SE 0.0012) to 0.808 (SE 0.0016); and PA rises from 0.728 (SE 0.0012) to 0.735 (SE 0.0023). A similar result was found on the word representation of highlights and held out participants, AUC rises from 0.791 (SE 0.0011) to 0.808 (SE 0.0018); and PA rises from 0.728 (SE 0.0016) to 0.738 (SE 0.0023).

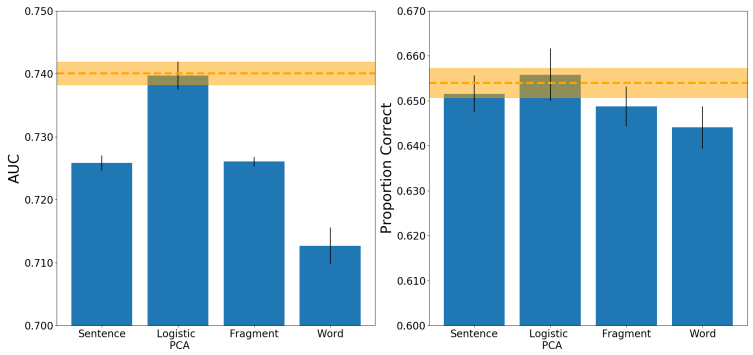


Figure 8: Results for the hold out questions cross validation split using the feature-based regression model for the task of predicting quiz score. The orange line is the baseline result.

# 5 Highlight Prediction

Having established that highlights have predictive utility on quiz performance, we turn to a related question: whether an individuals' highlights in one section of the text help to predict their highlights in other sections. For this investigation, we explore matrix factorization methods. Matrix factorization is a popular method for collaborative filtering—problems involving predicting missing features of one individual, based on population data. Examples of collaborative filtering problems include movie and product recommendations [18]; in these problems, the data set is a matrix of scores whose cells contain the rating of a given individual for a given item. In our case, the matrix, $\boldsymbol{H}$, is binary, where cell $h_{pi}$ indicates whether participant $p$ has highlighted text segment $i$. Matrix factorization methods decompose the matrix into a latent feature vector for each individual and a latent feature vector for each product, such that the value in a matrix cell depends on the compatibility of the corresponding latent feature vectors, which in essence assesses whether a given type of student tends to be interested enough to highlight a given type of material.

We use SPARFA-M [19] as the algorithm for matrix factorization. SPARFA-M infers $k$ latent concepts that are used to characterize each segment of text as well as each participant's interests. Every text segment $i$ is described as a $k$-element vector, $\boldsymbol{w}_i$, whose elements represent the contribution of each latent concept to the segment. Each participant $p$ is described by a $k$-element vector, $\boldsymbol{c}_p$, whose elements represent the participant's propensity to highlight each of the latent concepts. SPARFA-M models boolean random variables $X_{pi} \in \{0, 1\}$ that predict whether participant $p$ has highlighted text segment $i$, where

$$X_{pi} \sim \text{Bernoulli}(\text{logistic}(\boldsymbol{w}_i^T \boldsymbol{c}_p + \mu_i))$$

and $\mu_i$ represents the intrinsic propensity to highlight segment $i$. Following Lan et al. [19], we choose $k = 5$; as previous work also found, negligible increases in performance are obtained with larger values of $k$, at the cost of higher computation.

According to the model, the probability of an observed highlight $h_{pi}$ is

$$Pr(X_{pi} = h_{pi}) = \text{logistic}(\boldsymbol{w}_i^T \boldsymbol{c}_p + \mu_i)^{h_{pi}} + (1 - \text{logistic}(\boldsymbol{w}_i^T \boldsymbol{c}_p + \mu_i))^{1-h_{pi}}. \quad (3)$$

Given the highlight observation matrix $\boldsymbol{H}$, SPARFA-M performs maximum likelihood estimation with respect to $\boldsymbol{W} = [w_1\ w_2\ ...]$, $\boldsymbol{C} = [c_1\ c_2\ ...]$, and $\boldsymbol{\mu} = [\mu_1\ \mu_2\ ...]$. Sparsity constraints are incorporated into the estimation problem, yielding the regularized log likelihood:

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{C}, \boldsymbol{\mu}) = \sum_{i,p} \ln Pr(h_{pi} \mid \boldsymbol{W}, \boldsymbol{C}, \boldsymbol{\mu}) - \lambda_1 \sum_i ||\boldsymbol{w}_i||_1 - \frac{\lambda_2}{2} \sum_i ||\boldsymbol{w}_i||_2^2 - \frac{\lambda_3}{2} ||\boldsymbol{C}||_F^2,$$

where $\lambda_1$, $\lambda_2$, and $\lambda_3$ are regularization coefficients, whose values are selected using 3-fold cross validation on the training set. To prevent overfitting and improve identifiability, SPARFA-M has three key assumptions: (1) the number of latent concepts $k$ is small relative to the number of learners and questions, (2) $\boldsymbol{W}$ is sparse, and (3) the entries of $\boldsymbol{W}$ are non-negative. Optimization is performed via the FISTA framework [4].

## 5.1 Simulation Results

To assess highlight prediction, we performed ten-fold cross validation to split participants into training and evaluation sets. All highlighting data from participants in the training set were used for model training. For each participant in the evaluation set, we predict the word representation of the highlights in one passage conditioned on the highlights in the remaining two passages. The passage used for prediction was chosen at random; the other two passages provide context for the prediction. Simulations for the other types of representations (i.e. sentences, fragments, and LPCA) are excluded from this discussion because they obtained similiar performance to the word representation and these other representations can be derived from the word representation. We compare SPARFA-M to a baseline model whose prediction is simply the mean proportion of participants in the training set who highlight a given feature (word).

SPARFA-M outperforms the baseline model on predicting highlighting patterns (Table 2), indicating that the highlights a participant makes in several passages can be useful for determining deviations from population behavior in a third passage. The discrepancy between AUC and prediction accuracy can be explained by the fact that accuracy is sensitive to the decision criterion, and the two models are operating with different criteria. To illustrate this fact, we changed the decision criterion for prediction accuracy from 0.5 to 0.2, SPARFA-M obtained an accuracy of 0.688 (SEM 0.00708) which handily beat the baseline with an accuracy of 0.506 (SEM 0.00686). Because AUC provides a measure of discriminability insensitive to the decision criterion, we argue that AUC is more meaningful as a performance measure.

Table 2: Predictions of highlights in one passage given the highlights in the other two passages.

| | AUC | Accuracy |
|---|---|---|
| SPARFA-M | 0.765 (0.0101) | 0.742 (0.00788) |
| Baseline | 0.683 (0.00754) | 0.730 (0.00647) |

# 6 General Discussion

We described an experiment in which highlights are collected as participants read three passages from a biology textbook. Participants are then given the opportunity to review the passages and their highlights, and then take a quiz generated from the factual material from the text. We found evidence that the pattern of highlights an individual makes systematically varies between individuals, and that the pattern of highlights provides leverage in both predicting quiz performance and predicting which text will be highlighted in other passages. Highlights are a particularly promising data source given students natural inclination to highlight and their belief that highlighting is a valuable study strategy.

The improvement in prediction metrics with highlight data is small in magnitude but reliable. This finding is not surprising: we are using the highlighting choices as a proxy for the complex interpretative and memory processes a reader undergoes when exposed to novel material. Highlights provide a peek into these processes, but obviously not a complete record. As in many other big-data scenarios involving human learning and education, the hope is that many weak predictors can be identified and then combined to obtain stronger predictions. Digital textbooks provide other candidate predictors, including readers' gaze, page reading times, scrolling behavior, electronic notes and comments and questions, or even the geographic location and time at which the individual engages with the material.

A limitation of this study is the short span and limited content of our experiment. Only three passages are studied for under 30 minutes total. In contrast, a college course involves study of hundreds of pages over the span of a semester. We are optimistic that data collected from online learning platforms, such as Hypothes.is or Openstax—which offer the opportunity to observe student interactions with material over a broader range of content and longer period of time—will provide stronger insights into student knowledge and interests.

The modeling approach we have taken in this article will need to be extended as richer and larger data sources are obtained. First, we will need to advance from the simple regression and latent-variable models examined here to deep-learning models, whose complexity will be warranted given sufficient data. Second, we will need to give more consideration to the representation of highlights. Our passages were short enough that we could encode highlights by the specific position in the text they occupied. This encoding allowed us to sidestep issues of natural language content. For a larger (and evolving) text base, position-specific encodings will infeasible and would sacrifice the information content of the text itself. Consequently, we expect word and sentence embedding methods [10, 33] to be useful as we scale up the approach. This approach has the additional benefit of being robust to updates of a text; whereas traditional publishing methods produce infrequent updates, web-based publishing permits rapid and continual updating, and it is therefore prudent to merge highlighting data from different versions of a publication.

The finding that highlights have predictive value can lead to a variety of improvements to electronic textbooks:

- If highlights provide an early indication that a student is having difficulty comprehending or will have difficulty remembering material, early interventions can be performed to remedy the situation.
- Students whose highlights predict poor performance can be given training to improve highlighting strategies which in turn can improve their performance [22]. Even if tutorial advice was not provided to teach highlighting effectiveness, a pure data-driven approach is feasible: the predictive model could be inverted to determine highlighting patterns predictive of the best performance, and showing students highlights from an informed instructor have proven beneficial [15, 23, 29].
- Highlights predict not only performance, but also student focus and interest. To the extent that highlights can predict what will be highlighted in the future, recommender systems can be constructed to guide students to material likely to be of interest. Another approach to this same end is to cluster students by latent interests, as manifested in their highlighting patterns. We did not explore this clustering in the present article, but it is an interesting avenue for future research.

In conclusion, we have shown that highlights are a promising data source for assisting in early interventions and an important future work is to investigate whether other types of annotations, such as notes in the margin, also predict comprehension and affective behavior.

# 7 Acknowledgments

This research was supported by the National Science Foundation award EHR-1631428.

# 8 REFERENCES

[1] Steven P Abney. 1991. Parsing by chunks. In *Principle-based parsing*. Springer, Dordrecht, 257–278.

[2] Joseph Bafumi, Andrew Gelman, David K Park, and Noah Kaplan. 2005. Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis* 13, 2 (2005), 171–187.

[3] Frank B Baker and Seock-Ho Kim. 2004. *Item response theory: Parameter estimation techniques*. CRC Press, Boca Raton, Florida.

[4] Amir Beck and Marc Teboulle. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* 2, 1 (2009), 183–202.

[5] Kenneth E Bell and John E Limber. 2009. Reading skill, textbook marking, and course performance. *Literacy research and instruction* 49, 1 (2009), 56–67.

[6] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. (2017).

[7] Michael Collins, S. Dasgupta, and Robert E Schapire. 2002. A Generalization of Principal Components Analysis to the Exponential Family. In *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani (Eds.). MIT Press, Cambridge, MA, USA, 617–624. http://papers.nips.cc/paper/2078-a-generalization-of-principal-components-analysis-to-the-exponential-family.pdf

[8] Ryan SJ d Baker, Albert T Corbett, and Vincent Aleven. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Intelligent Tutoring Systems*. Springer, Springer, Berlin, Heidelberg, 406–415.

[9] Laurent Denoue and Laurence Vignollet. 2000. An Annotation Tool for Web Browsers and Its Applications to Information Retrieval. In *Content-Based Multimedia Information Access - Volume 1 (RIAO '00)*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, Paris, France, France, 180–195. http://dl.acm.org/citation.cfm?id=2835865.2835885

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. (2018).

[11] John Dunlosky, Katherine A Rawson, Elizabeth J Marsh, Mitchell J Nathan, and Daniel T Willingham. 2013. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest* 14, 1 (2013), 4–58.

[12] Rosta Farzan and Peter Brusilovsky. 2008. AnnotatEd: A social navigation and annotation service for web-based educational resources. *New Review of Hypermedia and Multimedia* 14, 1 (2008), 3–32.

[13] Robert L Fowler and Anne S Barker. 1974. Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology* 59, 3 (1974), 358.

[14] Zeno Gantner, Lucas Drumond, Christoph Freudenthaler, Steffen Rendle, and Lars Schmidt-Thieme. 2010. Learning Attribute-to-Feature Mappings for Cold-Start Recommendations. (2010), 176-185 pages.

[15] James Hartley, Sally Bartlett, and Alan Branthwaite. 1980. Underlining can make a difference—sometimes. *The Journal of Educational Research* 73, 4 (1980), 218–224.

[16] Peter W Hoon. 1974. Efficacy of three common study methods. *Psychological Reports* 35, 3 (1974), 1057–1058.

[17] Peter Idstein and Joseph R Jenkins. 1972. Underlining versus repetitive reading. *The Journal of Educational Research* 65, 7 (1972), 321–323.

[18] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. (2009), 30–37 pages.

[19] Andrew S Lan, Andrew E Waters, Christoph Studer, and Richard G Baraniuk. 2014. Sparse factor analysis for learning and content analytics. *The Journal of Machine Learning Research* 15, 1 (2014), 1959–2008.

[20] Andrew J Landgraf and Yoonkyung Lee. 2015. Dimensionality reduction for binary data through the projection of natural parameters. (2015).

[21] Seokho Lee, Jianhua Z Huang, and Jianhua Hu. 2010. Sparse logistic principal components analysis for binary data. *The annals of applied statistics* 4, 3 (2010), 1579.

[22] Detlev Leutner, Claudia Leopold, Den Elzen-Rump, et al. 2007. Self-regulated learning with a text-highlighting strategy: A training experiment. *Zeitschrift für Psychologie/Journal of Psychology* 215, 3 (2007), 174.

[23] Elizabeth Pugzles Lorch, Madeline A Klusewitz, et al. 1995. Effects of typographical cues on reading and recall of text. *Contemporary educational psychology* 20, 1 (1995), 51–64.

[24] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. (2014), 55–60 pages.

[25] Michael EJ Masson and Geoffrey R Loftus. 2003. Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 57, 3 (2003), 203.

[26] Aditya Krishna Menon and Charles Elkan. 2011. Link Prediction via Matrix Factorization. In *Machine Learning and Knowledge Discovery in Databases*, Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis (Eds.). Springer, Berlin, Heidelberg, 437–452.

[27] Caitlin Mills, Art Graesser, Evan F Risko, and Sidney K D'Mello. 2017. Cognitive coupling during reading. *Journal of Experimental Psychology: General* 146, 6 (2017), 872.

[28] Toshiya Miyatsu, Khuyen Nguyen, and Mark A McDaniel. 2018. Five Popular Study Strategies: Their Pitfalls and Optimal Implementations. *Perspectives on Psychological Science* 13, 3 (2018), 390–407.

[29] Sherrie L Nist and Mark C Hogrebe. 1987. The role of underlining and annotating in remembering textual information. *Literacy Research and Instruction* 27, 1 (1987), 12–25.

[30] Sherrie L Nist and Katie Kirby. 1989. The text marking patterns of college students. *Reading psychology: An international quarterly* 10, 4 (1989), 321–338.

[31] Philip I. Pavlik, Hao Cen, and Kenneth R. Koedinger. 2009. Performance Factors Analysis –A New Alternative to Knowledge Tracing. (2009), 8 pages. `http://dl.acm.org/citation.cfm?id=1659450.1659529`

[32] Radek Pelánek. 2015. Metrics for evaluation of student models. *Journal of Educational Data Mining* 7, 2 (2015), 1–19.

[33] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. Association for Computational Linguistics, Baltimore, Maryland, 1532–1543.

[34] Sarah E Peterson. 1991. The cognitive functions of underlining as a study technique. *Literacy Research and Instruction* 31, 2 (1991), 49–56.

[35] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep Knowledge Tracing. (2015), 505–513 pages. `http://papers.nips.cc/paper/5654-deep-knowledge-tracing.pdf`

[36] George Rasch. 1980. Probablistic Models for Some Intelligence and Attainment Tests. (1980).

[37] John P Rickards and Gerald J August. 1975. Generative underlining strategies in prose recall. *Journal of Educational Psychology* 67, 6 (1975), 860.

[38] Connie Rye, Robert Wise, Vladamir Jurukovski, Jung Desaix, and Yael Avissar. 2016. Biology. (2016).

[39] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. 2002. Methods and Metrics for Cold-start Recommendations. (2002), 8 pages. `https://doi.org/10.1145/564376.564421`

[40] Yue Shi, Martha Larson, and Alan Hanjalic. 2014. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)* 47, 1 (2014), 3.

[41] Keith A Wollen, Robert S Cone, Jerry C Britcher, and Karen M Mindemann. 1985. The effect of instructional sets upon the apportionment of study time to individual lines of text. (1985), 15 pages.

[42] Carole L Yue, Benjamin C Storm, Nate Kornell, and Elizabeth Ligon Bjork. 2015. Highlighting and its relation to distributed study and students' metacognitive beliefs. *Educational Psychology Review* 27, 1 (2015), 69–78.

[43] Liang Zhang, Xiaolu Xiong, Siyuan Zhao, Anthony Botelho, and Neil T. Heffernan. 2017. Incorporating Rich Features into Deep Knowledge Tracing. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale (L@S '17)*. ACM, New York, NY, USA, 169–172. `https://doi.org/10.1145/3051457.3053976`

# 9 Appendix

## Natural Language Processing

One may view the approach used to generate the fragment representation as a coarse version of chunking[1]. Chunking can be thought of as breaking up a sentence into it's constituents based on how it would be read. As an example consider the following sentence taken from [1]:

[I begin] [with an intuition]: [when I read] [a sentence], [I read it] [a chunk] [at a time].

Parsing textbooks at such a fine granularity such as this, would require enormous amounts of data to be able to constrain all the free parameters. Therefore we seek a representation that can achieve similar size and performance to the fragment representation. Ideally the previously parsed sentence would look like the one below:

[I begin with an intuition]: [when I read a sentence], [I read it a chunk at a time].

To obtain parses such as the one above, we generated constituency parse trees (using the Stanford CoreNLP parser [24]) for each sentence in the corpus and then descended the tree top-down, until we found a depth where there were two or more subtrees that contained all of the words in the sentence. These subtrees were defined as the features and thus each word in their respective subtree was assigned to the feature number of the subtree. An example parsed sentence can be found in Fig. 9. Compare the parse in Fig. 9 to the fragment parse of the same sentence:

**Constituency Parse:**
[Because people do not normally eat from cars or carpets],[these items][do not require the same level of cleanliness that silverware does].

**Fragment Parse:**
[Because people do not normally eat from cars or carpets],[these items do not require the same level of cleanliness that silverware does].

This scheme produced 340 features. Although the above example seems encouraging, this approach did not perform well as compared to the other representations.

In addition, we explored dependency parse trees with similar algorithms, however the performance was sub-par as compared to the constituency parse tree extractions, thus we exclude further discussion

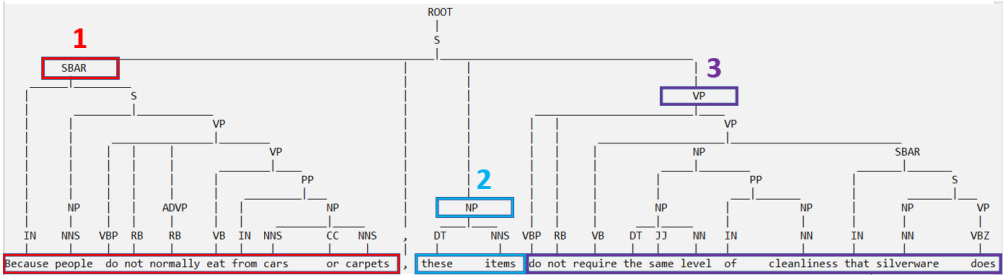


Figure 9: An example constituency parse of a sentence from the corpus with each color denoting a feature and the words which belong to the feature.