COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 44 (2020) e12901

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12901

Highlights as an Early Predictor of Student Comprehension and Interests

Adam Winchell,[a] Andrew Lan,[b] Michael Mozer[a,c]

[a]*Department of Computer Science, University of Colorado Boulder*
[b]*College of Information and Computer Sciences, University of Massachusetts Amherst*
[c]*Google Research*

Received 24 April 2019; received in revised form 6 August 2020; accepted 18 August 2020

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 her mental state—her level of comprehension, grasp of the key ideas, reading goals, and so on. We explore this hypothesis via an experiment in which 400 participants read three 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. Nonetheless, highlighting patterns may allow us to infer reader comprehension and interests. Using multiple representations of the highlighting patterns, we built probabilistic models to predict quiz performance and matrix factorization models to predict what content would be highlighted in one passage from highlights in other passages. We find that quiz score prediction accuracy reliably improves with the inclusion of highlighting data (by about 1%–2%), 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. Furthermore, an individual's highlighting pattern is informative of what she highlights elsewhere. Our long-term goal is to design digital textbooks that serve not only as conduits of information into the reader's mind but also allow us to draw inferences about the reader at a point where interventions may increase the effectiveness of the material.

Keywords: Bayesian modeling; Factor analysis; Intelligent textbooks; Learning analytics; Reading comprehension; Student modeling

Correspondence should be sent to Adam Winchell, University of Colorado, 430 UCB, 1111 Engineering Dr, Boulder, CO 80309. Email: adwi8859@colorado.edu

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 take 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 students are assigned a reading from a printed textbook and then take a quiz days later in class, the opportunity to perform prompt interventions has been lost. With the advent of electronic texts, data can be collected from students during their initial exposure to unfamiliar reading material, and if knowledge state can be inferred from these data, early interventions can be performed. To model engagement and comprehension, an obvious source of information is the gaze pattern of a student reading a textbook (Mills, Graesser, Risko, & D'Mello, 2017). However, reliable gaze data are quite difficult to collect in a naturalistic setting. Fortunately, explicit behavioral measures are often available: Given students' proclivity to highlight textbooks (Annis & Kent Davis, 1978; Bell & Limber, 2009; Kornell & Bjork, 2007; Lonka, Lindblom-Yl¨anne, & Maury, 1994), we can leverage highlights as a data source to draw inferences about individuals' attentional and knowledge states.[1]

Past research has examined the conditions under which highlighting impacts learning and memory. We review this literature in order to motivate the possibility that highlighting might reflect and influence cognitive states, and therefore be useful in predictive modeling.

1.1. The utility of highlighting

Highlighting may benefit a reader for two distinct reasons: First, it may encourage a deeper level of processing of the material (Craik & Lockhart, 1972); second, it may function as external memory (Faw & Waller, 1976), also known in the literature as a *storage function*, which can support subsequent study.

Fowler and Barker (1974) explored depth of processing by asking undergraduates to read articles from *Scientific American* in four highlighting conditions: *active*, in which students highlighted as much content found no difference between the techniques as they wanted to; *passive yoked*, in which students read marked texts that had been highlighted by yoked participants in the active condition; *passive expert-based*, in which students read marked texts that had been highlighted by the experimenters to reflect critical material; and *control*, in which students read articles without any highlights. The conjecture is that active highlighting might engage semantic analysis to select a subset of material, whereas passive highlighting would seem to benefit the student only by focusing attention relative to a condition in which no highlights appeared in the text. Thus, the conditions may affect the depth of processing that highlights induce.

No difference was found between conditions in overall exam score. However, if the sentences in the text that are critical for a given test item had been highlighted, that

item's performance benefits significantly. Specifically, active highlighting yields superior performance on these test items compared to passive-yoked highlights, and passive expert-based highlighting yields superior performance on these test items compared to the control condition. These findings provide weak evidence for the hypothesis that active highlighting engages a deeper level of processing than passive highlighting, which in turn is superior to not highlighting.

To push the depth-of-processing perspective further, one can require students to engage even more substantively with the material via a *constrained active* condition in which students are restricted to highlighting only one sentence per paragraph (Johnson, 1988; Rickards & August, 1975; Wollen, Cone, Britcher, & Mindemann, 1985). This constraint encourages students to identify the most important idea in a paragraph. Rickards and August (1975) investigated the difference between three conditions in which participants were asked to make one highlight per paragraph, either the sentence they deemed most important, least important, or any arbitrary sentence. Highlighting the most important sentence yields superior test performance to highlighting the least important sentence or to not highlighting at all. However, students who can highlight any single sentence without restrictions performed best among all groups and recalled significantly more incidental material than the group who highlighted the most important sentence per paragraph. The authors argue that the task demand for the group that highlighted the most important sentence per paragraph may have mitigated against learning the less important material. Nonetheless, imposing a constraint on the quantity of text highlighted may encourage "active, constructive learning" (Rickards & August, 1975, p. 865), and thereby boosts memory retention.

By definition, highlighting obviates the need for students to explicitly synthesize and reconstruct content. Consequently, the generation effect (Slamecka & Graf, 1978) might suggest that highlighting could harm performance by enabling readers to treat the superficial act of highlighting as sufficient engagement with the material. However, active highlighting does demand selection and discrimination of content, and thus might serve as a closer proxy to generation than a pure reading strategy. Consistent with the notion that active highlighting shares some of the benefits of generation, the effects of each are more potent when students have some prior knowledge of the material (for the generation effect, see Lutz, Briggs, & Cain, 2003; McNamara, 1995; for highlighting, see Blanchard & Mikkelson, 1987; Klare, Mabry, & Gustafson, 1955; Wollen et al., 1985).

Beyond the depth-of-processing argument, a distinct potential benefit of highlighting is its use as external memory. Highlighting can be viewed as facilitating the von Restorff effect (Cashen & Leicht, 1970; Fowler & Barker, 1974; Nist & Hogrebe, 1987; Wallace, 1965; Yue, Storm, Kornell, & Bjork, 2015), wherein the highlighted items stand out and thus are more memorable. Moreover, the salience of these items makes them easier to review. This phenomenon can be observed in the studies obtaining improved recall for passive highlighting or experimenter-provided highlights (Cashen & Leicht, 1970; Crouse & Idstein, 1972; Hartley, Bartlett, & Branthwaite, 1980).

Although highlighting has potential to benefit learners, some analyses find no benefit or detrimental effects of highlighting. For instance, Peterson (1991) found a disadvantage

for inference-based questions when students highlight, and Dunlosky, Rawson, Marsh, Nathan, and Willingham (2013) argue that highlighting may be more beneficial for calling attention to individual concepts and facts rather than the connections among them; however, Ben-Yehudah and Eshet-Alkalai (2018) found a benefit of highlighting for inference-based as well as fact-based questions. Ben-Yehudah and Eshet-Alkalai (2018) also examined the effect of presentation medium and found that highlighting improves exam scores relative to a no-highlight control for print media, but not for digital media. Possibly, depth of processing is limited for digital media due to demands on motor control to point and highlight with a cursor. Other researchers have argued that highlighting is ineffective because students often do not know how to highlight and thus the activity degenerates to a mechanism for tracking position in the text which requires only superficial processing (Hoon, 1974; Idstein & Jenkins, 1972; Nist & Kirby, 1989); this concern may be exacerbated for low-skilled readers who are less capable of marking important content (Bell & Limber, 2009). It is therefore not entirely surprising that recent reviews of the literature have reached diverging conclusions about the benefits of highlighting (contrast Dunlosky et al., 2013, and Miyatsu, Nguyen, & McDaniel, 2018).

1.2. Our contributions

Nearly all past work has examined highlighting at a condition level, asking whether one variety of highlighting yields enhanced student learning over another variety. When an individual student's specific highlights are considered, they are typically cast in terms of summary statistics such as the proportion of sentences highlighted, the number of main concepts highlighted, and the criticality of sentences highlighted (Blanchard & Mikkelson, 1987; Fowler & Barker, 1974; Idstein & Jenkins, 1972; Johnson, 1988; Nist & Kirby, 1989; Rickards & August, 1975; Wollen et al., 1985; Yue et al., 2015). In this article, we focus on individual students, asking whether the specific pattern of highlights that one student makes can predict whether he or she will learn the material better than another student with a different pattern of highlights. We thus move from the domain of cognitive psychology to the domain of data mining.

Past work has also emphasized dependent measures based on overall performance, either recall or the cloze procedure (Annis & Kent Davis, 1978; Blanchard & Mikkelson, 1987; Cashen & Leicht, 1970; Hartley et al., 1980; Idstein & Jenkins, 1972; Johnson, 1988; Leutner, Leopold, & Elzen-Rump, 2007; Wollen et al., 1985). Relatively little research has focused on accuracy on specific questions (Fowler & Barker, 1974; Nist & Hogrebe, 1987; Peterson, 1991; Yue et al., 2015). In this article, we construct models parameterized for specific questions and conditioned on the detailed pattern of highlights, offering in principle a finer granularity of prediction.

The emphasis of past work has been on understanding when highlighting is beneficial and what form of highlighting is beneficial. Our data mining perspective suggests that another interesting question to ask is whether one can predict an individual's specific pattern of highlights on new material given the individual's previous highlights. Such

predictions might be useful for guiding a student to material that interests them or detecting when students are mind wandering by the deviation from their predicted behavior.

To explore the value of highlights as a data source, we conducted an experiment in which participants read and highlighted sections of an electronic textbook, reviewed the material with highlights, and then took a delayed quiz. Observed highlights were used to predict quiz scores using feature-based regression models that extend item-response theory (Rasch, 1980). The pattern of highlights from two passages was also used to predict the highlights in a third passage using matrix factorization models from the collaborative filtering literature (Shi, Larson, & Hanjalic, 2014).

To summarize our results, highlights improve predictions of quiz performance for held-out participants or held-out participant questions (i.e., holding out a random subset of questions from each participant), but not for held-out questions. The predictive power of highlights is modest in magnitude, but highlights do offer information about the participant's knowledge state, above and beyond the participant's mean ability and a question's mean difficulty. In exploring various representations of the highlights, we found that coding them in terms of word primitives, that is, as a feature vector indicating which words are highlighted, achieves the best predictions. We also found that an individual's highlighting pattern informs predictions of what he or she would highlight elsewhere.

2. Methodology

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 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 rereading of the text would not be feasible.

2.1. 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 were discarded. The experiment took 25–30 min to complete. To incentivize participants, they 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 that they were familiar with the experiment material (despite having indicated no familiarity in advance of the experiment).

2.2. Materials

Three passages were selected from the Openstax *Biology* textbook (Rye, Wise, Jurukovski, Desaix, & Avissar, 2016). Biology was used as a domain for several reasons. First, the foundations of biology are a set of concepts and facts upon which scientific understanding is built; improving retention of these fundamental concepts and facts is critical to mastery of the field. Second, the Openstax biology textbook has been widely adopted, and we hope to build on the current study to a follow-up study using actual students in college-level courses. The choice of text is likely not critical: Dunlosky et al. (2013) note that a variety of content domains have been used to study highlighting, from aerodynamics to ancient Greek schools, and the pattern of results has been similar regardless of the domain.

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 12 questions were transformed into 12 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* in the range 0–1 reflecting 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.

2.3. 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 and the makeup of the quiz, and they were encouraged to highlight by being told that the highlights would be available during the 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 3 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 5 min; in Condition B, each passage was displayed for 6 min; the increase in time between Conditions was to alleviate participant concerns of feeling rushed during the reading of the more technical passage as indicated by emails detailing these concerns. 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 would cause the full word to be highlighted. Participants could unhighlight by selecting a previously highlighted sequence of words. 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 it is not allowed to 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 1 min. 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 was cleared and a message was displayed describing the next step of the experiment. Throughout the course of the experiment, a progress bar was displayed at the bottom of the screen that indicated the 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. (The reason for blocking questions was to better control the time between reading of the passage and the quiz.) At the end of the quiz phase, we asked participants in Condition B whether in retrospect the material was familiar to them. Three participants 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.

3. Results

Overall, participants correctly answered 35% of fill-in-the-blank (FIB) questions and 64% of MC questions. Answering the FIB version of a question boosted the accuracy of the MC version of the questions to 83% (Table 1). In the three sections that follow, we (a) perform traditional hypothesis testing to explore the relationship between summary statistics of the highlights and quiz score, (b) construct models that use the specific pattern of highlights to predict quiz score, and (c) construct models that use the specific pattern of highlights in two passages to predict highlights in a third passage.

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	0.30	0.35
FIB Correct	0.06	0.29

3.1. Hypothesis testing

We conducted a two-way ANOVA to compare the effect of experiment Condition (A vs. B) and passage number (1, 2, 3) on quiz score, with participants as the random factor. A main effect of experiment Condition is observed ($F(1, 389) = 7.38$, $p = .007$), with participants in Condition A performing better (50.8% vs. 46.3%); we conjecture the reason is that Condition B ensures there is some lag between reviewing a passage and being asked a question on that passage, preventing access to quiz answers from working memory. A main effect of passage number is observed ($F(2, 778) = 53.6$, $p < .001$), with the questions associated with one passage being harder than those of the other two (42.0% vs. 52.2% vs. 51.3%). Importantly, no interaction is observed between version and passage ($F(2, 778) = 1.45$, $p = .24$); therefore, in all subsequent analyses, we combine data from the two conditions.

We next examine the relationship between summary statistics of highlighting and overall 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 appears to be a mixture of a unimodal distribution and an impulse function at 0 (individuals who did not highlight). The normalized quiz score, shown along the right margin, is unimodal with a mean of 0.49. Although the scatter plot suggests no strong functional relationship between the amount of text highlighted and quiz performance, the correlation coefficient is 0.17 ($p < .001$). This correlation drops to 0.079 ($p = .15$) when participants who did not highlight are removed from the analysis.

Because each question in our quiz is based on a specific sentence in the text passages, we can define the *critical sentence* required to answer a given quiz question. Following Yue et al. (2015), we investigated whether highlighting the critical sentence is related to performance on the corresponding quiz question. Yue et al. calculated a *highlighting efficiency* score for each participant, defined as the ratio of highlighted critical sentences to the total number of highlighted sentences. The correlation between highlighting efficiency and the normalized quiz score is 0.12 ($p = .028$). Thus, neither the absolute nor relative number of highlights is a robust predictor of quiz performance.

Turning to an analysis of specific quiz questions, we examine the relationship between whether or not a critical sentence is highlighted and whether or not the corresponding question is answered correctly. Analyzing the FIB and MC questions separately, a two-tailed matched sample *t*-test with participants as the random factor indicates accuracy is

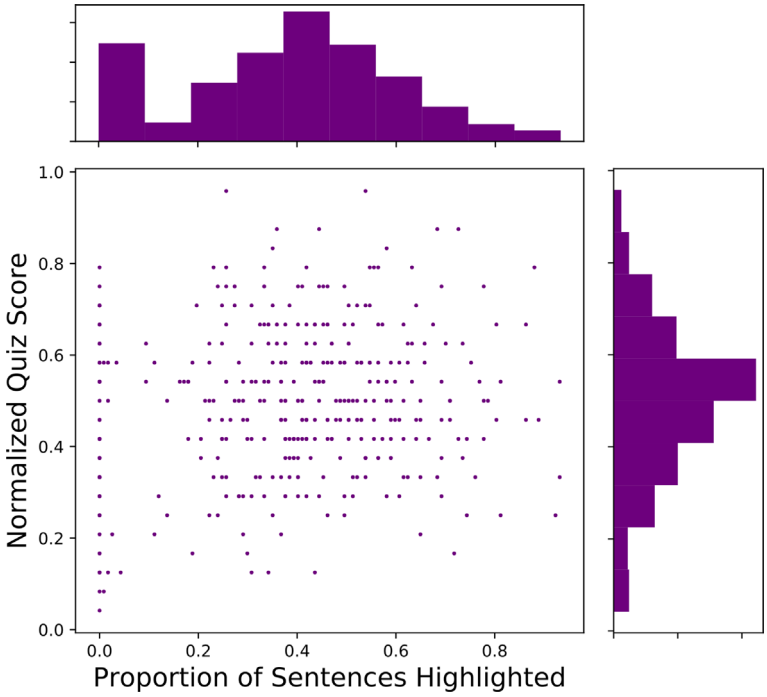

Fig. 1. Scatter plot of proportion of sentences highlighted versus normalized quiz score for each participant. The marginal distributions are shown above and to the right of the scatter plot.

significantly higher for individuals who highlight the critical sentence than for those who did not, both for FIB questions (37.3% vs. 29.8%, $t(296) = 3.23$, $p = .001$, $d = 0.27$) and MC questions (68.2% vs. 62.0%, $t(296) = 4.14$, $p < .001$, $d = 0.70$). Participants who highlighted none or all of the critical sentences had to be excluded from this 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 ($F(2, 198) = 3.24$, $p = .041$), 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 ($F(2, 198) = 4.93$, $p = .0081$). This positive relationship between survey responses and quiz score contrasts with a negative relationship observed by Yue et al. (2015). Their experiment was remarkably similar in structure to ours, and it 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.

3.2. Predicting quiz score

We next examine the specific patterns of highlights and their relationship to performance. There is diversity in the manner in which individuals highlight, as illustrated in Fig. 2, which shows highlights produced by three participants for one specific paragraph of text. Highlights demarcate blocks of text ranging from single words to phrases to complete sentences. To distinguish among these highlighting patterns, we must specify an encoding of the highlighting data.

3.2.1. Highlight encodings

In all analyses, we ignore the time course and sequence of text selections and deletions that the participant made. Instead, we consider only the final highlighted state of each passage. We can parse the highlights at the granularity of a *sentence* and count a sentence as highlighted if any word in the sentence is highlighted. (We define sentences as delineated by periods, exclamation marks, and question marks.) We can also parse highlights at the granularity of individual *words*. Intermediate between sentences and words is the level of *fragments*, which are phrases delineated by commas, semicolons, and colons. We subjectively excluded some commas as segment boundaries, for example, commas that delineate lists of items. Fig. 3 gives an example of our fragment partitioning. A fragment is counted as highlighted if any word in the fragment is highlighted. The three passages contain a total of 2,291 word tokens, 235 fragments, and 117 sentences.

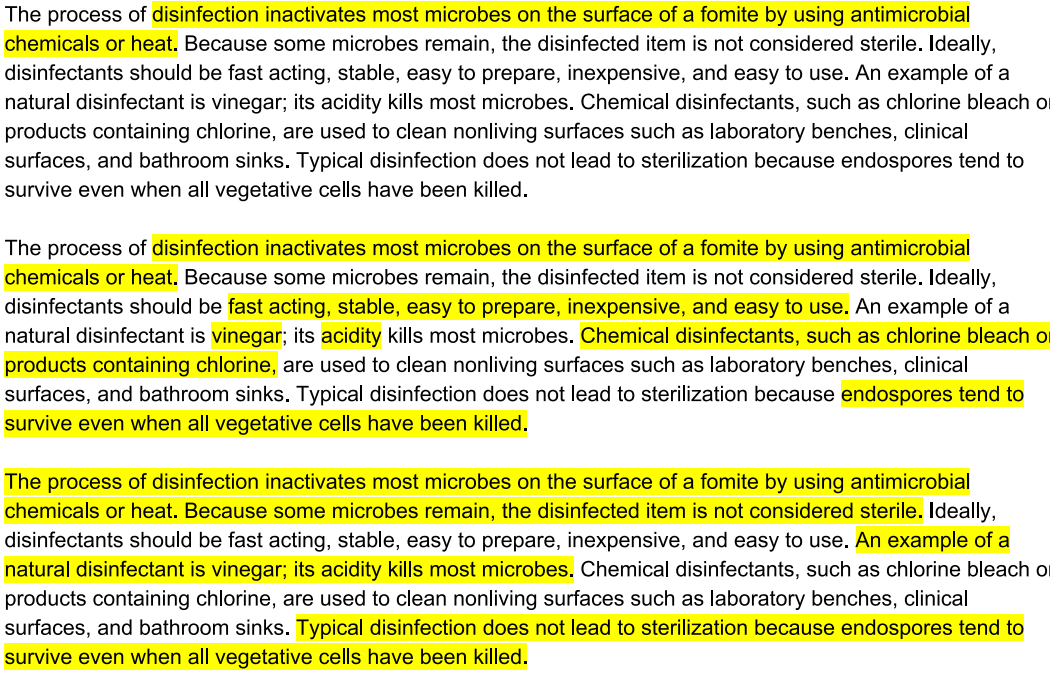

Fig. 2. A paragraph of text as highlighted by three randomly selected participants.

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 fragment representation was intended as a compromise, capturing detail while not requiring too many free parameters. In contrast to our manually segmented fragment representation, we explored two additional intermediate-level representations, both derived automatically: logistic principal components analysis (LPCA; Collins, Dasgupta, & Schapire, 2002; Lee, Huang, & Hu, 2010) and a traditional natural-language processing methodology for segmentation into fragments.

LPCA is an unsupervised method for binary data. Traditional principal components analysis (PCA) seeks to find a small set of orthogonal components that capture the variance in 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 2,291 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 of Landgraf and Lee (2015), we performed fivefold cross-validation to tune a hyperparameter.[2]

We also considered a representation based on a traditional natural-language processing methodology, *constituency parsing*, which breaks a sentence into its constituents or chunks, as in the following example (Abney, 1991):

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

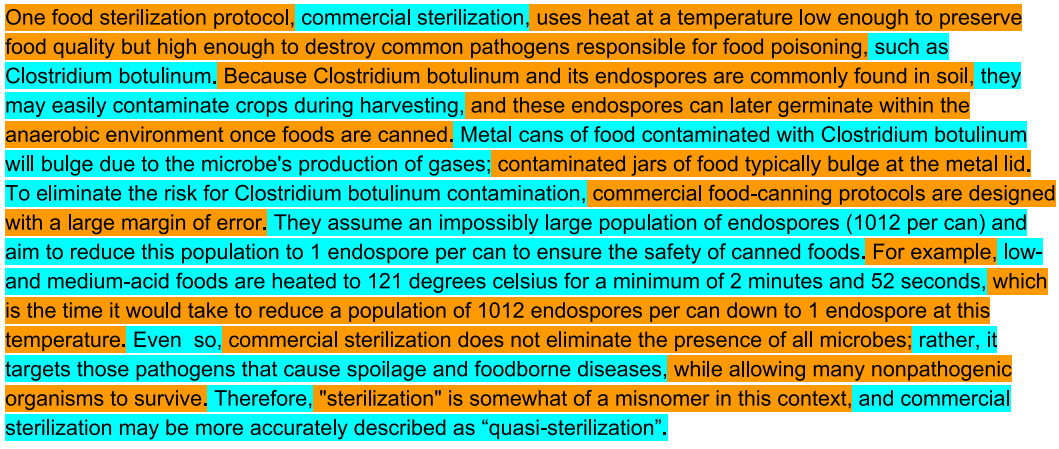

Fig. 3. Example of the fragment representation where the alternating colors signify the different fragments.

Although one could construct a highlighting representation based on such constituents, it would be much higher dimensionality than our fragment representation. To reduce the dimensionality, we used the following procedure. First, we generated constituency parse trees for each sentence in the corpus using the Stanford CoreNLP parser (Manning et al., 2014). Then, we descended the tree from the root until we identified a depth at which all the words in a sentence were contained in two or more subtrees. We created a feature for each subtree, yielding 340 features that broke up the passages in a manner similar to that achieved by the fragment representation. However, the performance of the automatically derived subtree representation yielded results that were significantly worse than those from the fragment representation. We thus excluded it from our representation. We also explored a representation based on dependency parse trees using the Stanford CoreNLP parser (Manning et al., 2014), but it performed even worse.

3.2.2. Quiz prediction models

We now turn to the prediction of quiz performance given an individual's highlights, as represented by one of the schemes described in the previous section. Following a long tradition in the educational data mining community, we use feature-based regression models. Feature-based regression models include performance factor analysis (PFA; Pavlik, Cen, & Koedinger, 2009) and deep knowledge tracing (Piech et al., 2015); these two approaches differ in that features are handcrafted in the former and learned from the data in the latter. In addition to features that encode past history of student performance, some models have incorporated side information, that is, information not directly related to the dependent measure being predicted. Examples of side information include viewing times and requests for hints (e.g., Baker, Corbett, & Aleven, 2008; Zhang, Xiong, Zhao, Botelho, & Heffernan, 2017). Highlighting patterns constitute a novel source of side information. In addition to observable features, some 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 (Baker & Kim, 2004), forms the backbone of our regression approach. Item-response theory combines side information and latent 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:

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

where αp denotes the latent ability of individual p, δ_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 vs. 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 ν_f and β_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 $\{\alpha, \delta, \nu, \beta\}$ 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 identifiability issues, ν_f and βe were used to identify the model (see Bafumi, Gelman, Park, & Kaplan, 2005, for advice on dealing with model identification). All of the feature-based regression models were fit using STAN (Carpenter et al., 2017). We sample four MCMC chains each having 2,500 samples,[3] 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, h_p for participant p and an associated highlight coefficient vector ω_{c_i}, yielding:

$$\Pr(y_{pi} = 1) = \text{logistic}(\alpha_p - \delta_{c_i} + \nu_{f_i} + \beta_{e_p} + \omega_{c_i} h_p^T). \tag{2}$$

The hyperparameters of the baseline model are reused with priors on the highlighting coefficients being $\omega_{c_i} \sim N(\mu_\omega, \sigma_\omega)$.

3.2.3. Performance metrics

Following the guidance of Pelánek (2015), we evaluate models with two performance measures: area under the ROC curve (AUC) and prediction accuracy. AUC measures how well a model discriminates between classes. AUC typically lies in the range between 0.5 (no ability to discriminate correct from incorrect) and 1.0 (perfect discrimination). AUC was computed by merging all the predictions in the evaluation set. Prediction accuracy 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 value of highlights on quiz score 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, α_p and δ_{c_i} respectively, in Eq. 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 (Gantner, Drumond, Freudenthaler, Rendle, & Schmidt-Thieme, 2010; Schein, Popescul, Ungar, & Pennock, 2002). We use 10-fold splits for the held-out participants and participant questions, and 12-fold splits for the held-out questions.

3.2.4. Simulation results

Consider the model in Eq. 2 with a sentence representation of highlights. For each quiz question q, this model has a coefficient vector, ω_q, that indicates whether highlighting specific sentences increases or decreases the probability of correctly answering question q. Fig. 4 depicts these coefficients via an array with one row per question and one column per sentence in the text. The coloring indicates the sign and magnitude of the coefficient for the highlighting of a given sentence on a given question. The coefficients are normalized by question such that the magnitude of the largest is 1.0. These coefficients are obtained by averaging across cross-validation folds of models trained on subsets of participants. In each row, the black square indicates the critical sentence in the text that is sufficient to answer the corresponding quiz question.

Notably, the coefficients on the critical sentences are not the largest in each row, indicating that the model has identified stronger regularities than the straightforward rule that highlighting a sentence increases the probability of correctly answering questions about that sentence. The coefficient matrix seems to validate our modeling approach over previous analyses that have focused on whether the critical sentence is highlighted (Fowler & Barker, 1974). Although the coefficients reflect statistical regularities in our dataset, it is

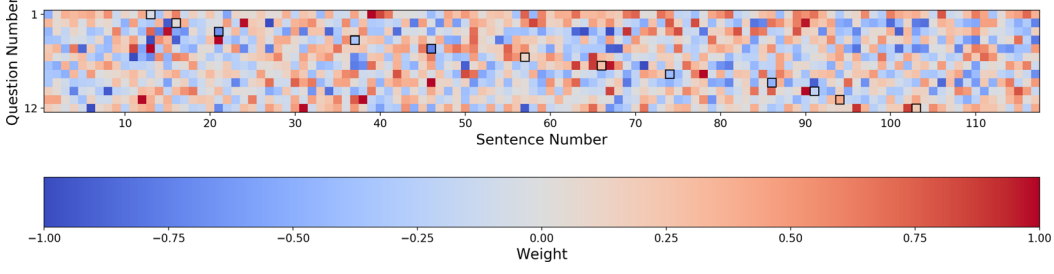

Fig. 4. Mean highlighting weights, ω_{c_i}, for models trained on the sentence representation. The weights are normalized by question such that the magnitude of the largest weight is 1.0. The black squares indicate which highlight features were used to generate each quiz question.

of course possible that they are not predictive of held-out data. To answer this question, we perform cross-validation.

Fig. 5 shows cross-validation performance of models trained on the four different highlight representations (see Section 3.2.1). 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 (Eq. 1), which does not utilize highlighting features. Shading represents ± 1 standard error of the mean.[4]

The four representations are ordered by granularity, that is, 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. 5, 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. 6) 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.

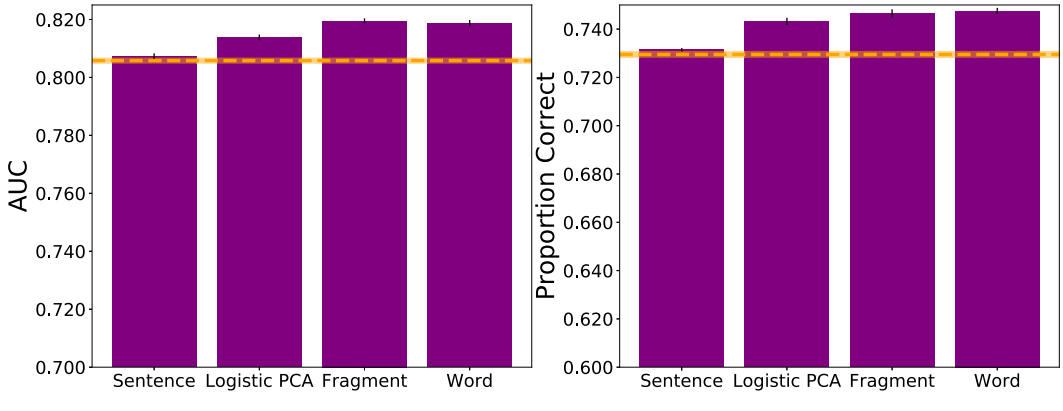

Fig. 5. Results for the hold-out participant questions cross-validation split using the feature-based regression model for the task of predicting quiz score. The orange line is the baseline result.

Consequently, participants' comprehension could be assessed online as they first engage with a text.

Prediction on unseen questions (Fig. 7) 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 reuse a set of (factual) questions from year to year, and these questions underlie the core material the instructor wishes students to master. 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 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, ω_{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).

3.3. Predicting pattern of highlighting

Having established that highlights have value in predicting quiz performance, we turn to a related question: whether an individual's highlights in one section of the text help to predict her highlights in other sections. Highlights indicate concepts that students believe are key, but they are also a proxy for what a student finds interesting and worthy of

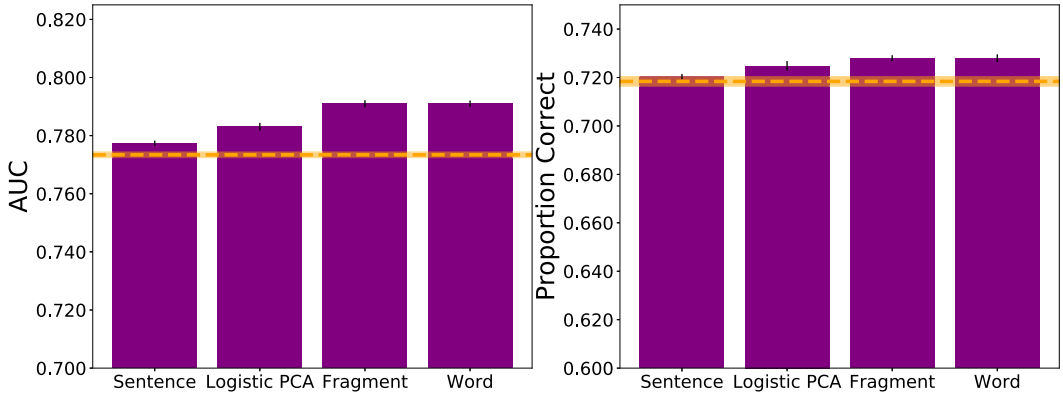

Fig. 6. Results for the hold-out participants cross-validation split using the feature-based regression model for the task of predicting quiz score. The orange line is the baseline result.

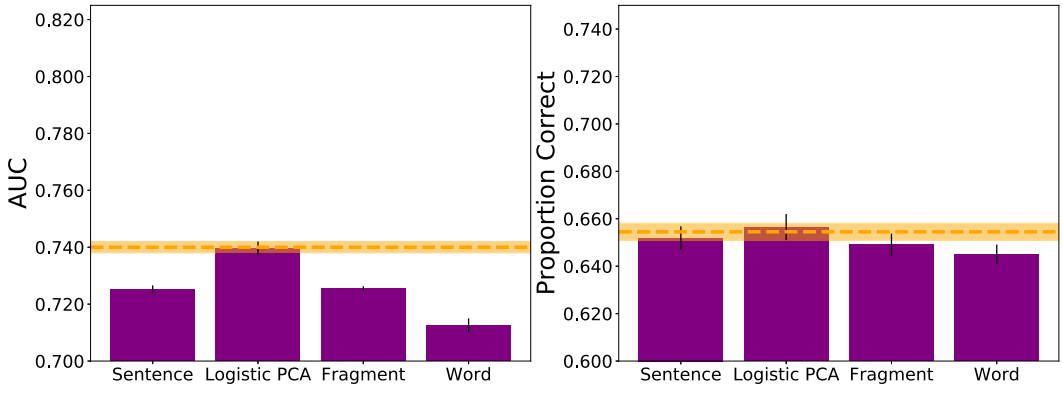

Fig. 7. Results for the hold-out questions cross-validation split using the feature-based regression model for the task of predicting quiz score. The orange line is the baseline result.

noting. Predicting highlights might therefore be useful for making recommendations to students concerning additional readings. Furthermore, discrepancies between predicted and actual highlights could possibly serve as an indication of mind wandering or other failures to engage with the material, which in turn could trigger educational interventions.

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 (Koren, Bell, & Volinsky, 2009); 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 segment, 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 bit of material.

We use SPARFA-M (Lan, Waters, Studer, & Baraniuk, 2014) 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, 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

$$X_{pi} \sim \text{Bernoulli}(\text{logistic}(w_i^T c_p + \mu_i)),$$

and μ_i represents the intrinsic propensity to highlight segment i. Following Lan et al. (2014), we choose $k = 5$; as previous work also found, negligible increases in performance are obtained with larger values of k, and they incur an increased computational cost.

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

$$\Pr(X_{pi} = h_{pi}) = \text{logistic}(w_i^T c_p + \mu_i)^{h_{pi}} + (1 - \text{logistic}(w_i^T c_p + \mu_i))^{1-h_{pi}}. \quad (3)$$

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

$$\mathscr{L}(\boldsymbol{W}, \boldsymbol{C}, \boldsymbol{\mu}) = \sum_{i,p} \ln \Pr(h_{pi} | \boldsymbol{W}, \boldsymbol{C}, \boldsymbol{\mu}) - \lambda_1 \sum_i ||\boldsymbol{\omega}_i||_1 - \frac{\lambda_2}{2} \sum_i ||\boldsymbol{\omega}_i||_2^2 - \frac{\lambda_3}{3} \sum_i ||\boldsymbol{C}||_F^2,$$

where λ_1, λ_2, and λ_3 are regularization coefficients, whose values are selected using three-fold cross-validation on the training set. To prevent overfitting and improve identifiability, SPARFA-M has three key assumptions: (a) the number of latent concepts k is small relative to the number of learners and questions, (b) $\boldsymbol{W}$ is sparse, and (c) the entries of $\boldsymbol{W}$ are non-negative. Optimization is performed via the FISTA framework (Beck & Teboulle, 2009).

3.3.1. Simulation results

To assess highlight prediction, we performed 10-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. 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).

As summarized in Table 2, SPARFA-M outperforms the baseline model on predicting highlighting patterns, indicating that the highlights a participant makes in several passages can be useful for determining deviations from population behavior in a third passage. These results are for the word representation of highlights. Representations based on sentences, fragments, and LPCA obtain similar performance.

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

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)

provides a measure of discriminability insensitive to the decision criterion, we argue that AUC is more meaningful as a performance measure for this task, which is echoed in the literature (Menon & Elkan, 2011).

Fig. 8 provides some insight into SPARFA-M's use of the latent concepts to predict the missing highlights. For each of the three passages and each of the five latent concepts, we computed the mean SPARFA-M word weighting for the given passage and concept.[5] The figure reveals that each passage has a strongly associated concept, and that concept is not associated with the other two passages. Thus, SPARFA-M has achieved a sort of segmentation by passage. We also examined the student × concept matrix, but we failed to identify any visualization or clustering that offered insight into the model's latent representations.

4. General discussion

We described an experiment in which participants read three passages from a biology textbook and were allowed to highlight as they read. Following the initial reading, participants had the opportunity to review the passages with highlights. Then, they took a quiz generated from the factual material from the text. We found that an individual's idiosyncratic pattern of highlights in a given section of text helps to predict the performance on quiz questions from this section, as well as to predict what words will be highlighted in other sections.

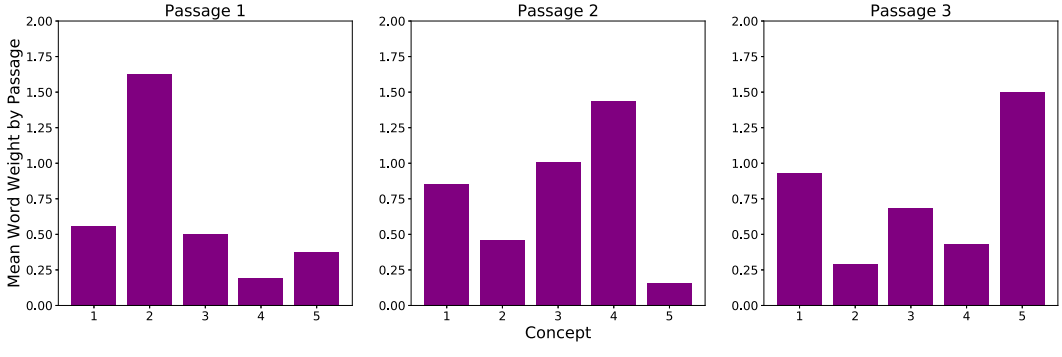

Fig. 8. Bar graphs of the mean word weighting in the given concept for each of the three passages taken from one fold of the cross-validation.

The improvement in prediction metrics that result from using highlighting 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. Whereas data collection from many sources requires effort or inconvenience on the part of the learner (e.g., surveys, EEG or skin response recording, gaze tracking), highlighting is natural to many students and thus imposes little burden. Fortunately, digital textbooks provide other lightweight data sources, including page reading times, scrolling behavior, electronic notes and questions, and the geolocation and time at which an individual engages with the material.

Although the focus of our work is on whether highlights are useful as a data source for modeling student comprehension, our data provide some evidence concerning theoretical issues, in particular, whether highlighting might engage a deeper form of processing of the material. We found a small but reliable benefit in performance for participants who highlight over those who do not. And we found that individuals who were more selective in their highlights—as measured by highlighting efficiency—showed a marginal benefit in performance. However, it is impossible to determine whether highlighting choices are the causal factor or merely correlated with a latent ability factor.

A limitation of this study is the short span and limited content of our experiment. Only three passages are studied for under 30 min total, and the lag between study and test is fairly brief. It is possible that the benefit of highlighting is fleeting: A student may remember the highlights he or she made 5 or 10 min earlier and may thus benefit from highlighting, but this benefit diminishes if the test is delayed by a week. We acknowledge this limitation, but argue that *indirect* benefits of highlighting may occur. For example, the external memory function of highlighting may lead to more consideration given to the highlighted paragraphs on delayed review. And even a short-lived memory boost from highlighting may facilitate understanding of subsequently learned material: To scaffold one concept upon another, the recently learned concepts must be accessible.

From a data mining perspective, the short span and limited content of our experiment impose a lower bound on what inferences may be drawn from highlights. Consider the data that could be collected from an online college course involving the study of hundreds of pages of text over the span of a semester. Online learning platforms, such as Hypothesis or Openstax, offer the opportunity to observe student interactions with a broader range of content over longer periods of time. The richer the data source, the more likely statistical methods are to find higher order regularities. This fact is in large part the explanation for the success of deep learning over the past decade.

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 be infeasible and would sacrifice the information content of the text itself. Consequently, we expect word and sentence embedding methods (Devlin, Chang, Lee, & Toutanova, 2018; Pennington, Socher, & Manning, 2014) 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.

To the extent that highlights are reliably predictive—of either learning or student interests—they might be leveraged to improve electronic textbooks in several ways.

1. If highlights provide a reliable indicator that a student is having difficulty comprehending or remembering material, immediate interventions might be performed to address the situation.
2. Students with highlighting patterns that predict poor performance or that do not match highlighting patterns of an instructor might be offered training to improve highlighting strategies (Leutner et al., 2007). Indeed, Miyatsu et al. (2018) and Yue et al. (2015) argue that training students to highlight appropriately is critical to obtaining value from the study strategy, and showing students highlights from an informed instructor has proven beneficial (Hartley et al., 1980; Lorch & Klusewitz, 1995; Nist & Hogrebe, 1987). Following training, data-driven approaches might be used to assess adherence to a strategy. Even if highlights from an informed instructor are not available, one could evaluate the success of a pure data-driven approach in which the predictive model is inverted to determine patterns predictive of best performance.
3. 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 might be constructed to guide students to material likely to be of interest. Kintsch (1980) proposed a distinction between cognitive and emotional interests. Whereas cognitive interests pertain to the key substantive content of a text, emotional interests capture attention even if they are peripheral to the main lessons. Readers are likely to be motivated and energized by emotional interests (Harp & Mayer, 1997). One intriguing direction for future research is to tease apart highlights that reflect these distinct interests. Another possibility is to cluster students by the latent interests manifested in their highlighting patterns, with the potential for forming work groups that either share interests or have complementary interests.

It might be considered ambitious to imagine that the modest predictive value we observe for highlights would be sufficient to reliably target students with an intervention. However, the subtle patterns we observe might provide sufficient evidence for a skilled human instructor to determine whether and what sort of intervention might benefit the student. And automatic interventions, even if poorly targeted, will provide additional feedback useful for improving comprehension models. For instance, if a student is given

an immediate question to bolster her comprehension, her answer provides a stronger indication of comprehension than does a delayed quiz.

In conclusion, we have shown that highlights are a viable data source for inferring the cognitive state of the reader. If we are able to obtain some signal in a laboratory experiment, we are optimistic that an even stronger signal may be obtained in an authentic learning scenario. Our current research is moving from the laboratory to an electronic textbook platform. Our future work will also investigate whether other types of annotations, such as notes in the margin, and implicit measures, such as page dwell times and scrolling behavior, can also be of assistance in the prediction of retention, comprehension, and engagement.

Acknowledgments

This research was supported by the National Science Foundation award EHR-1631428. The data for the experiment can be found at https://github.com/adam-winchell/Highlighting-Research.

Open Research badges

This article has earned Open Data badge. Data is available at https://github.com/adam-winchell/Highlighting-Research.

Notes

1. Underlining and highlighting are treated as equivalent techniques in the literature. For example, Fowler and Barker (1974) found no difference between the techniques.
2. The hyperparameter, m, restricted the logit of the observed probability to lie in $[-m, +m]$ rather than $[-\infty, +\infty]$. We considered $m \in [1, 10]$. Cross-validation selected $m = 7$.
3. The number of iterations was chosen based on models with the largest number of features—and hyperparameters—while maintaining convergence as determined by the split R statistic and Bayesian fraction of missing information (Betancourt, 2017).
4. The standard errors have been adjusted to remove variability arising from the random factor—the data split—while retaining variability across data splits in the relative performance of highlighting representations. We use the procedure suggested

by Masson and Loftus (2003), which results in corrected error bars whose overlap indicates whether performance differences are statistically meaningful.

5. Because the concept indices are arbitrary, the indices shift from one fold of cross-validation to the next. Thus, we show only one fold in the visualization. Others are similar in their interpretation.

References

Abney, S. P. (1991). Parsing by chunks. In R. C. Berwick, S. P. Abney, & C. Tenny (Eds.), *Principle-based parsing* (pp. 257–278). Dordrecht: Springer.

Annis, L., & Kent Davis, J. (1978). Study techniques and cognitive style: Their effect on recall and recognition. *The Journal of Educational Research*, *71*(3), 175–178.

Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, *13*(2), 171–187.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. Boca Raton, FL: CRC Press.

Baker, R. S., Corbett, A. T., & Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In B. P. Woolf, E. Aïmeur, & R. Nkambou (Eds.), *Intelligent tutoring systems*, Lecture Notes in Computer Science, *5091*, 406–415). Berlin: Springer.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, *2*(1), 183–202.

Bell, K. E., & Limber, J. E. (2009). Reading skill, textbook marking, and course performance. *Literacy Research and Instruction*, *49*(1), 56–67.

Ben-Yehudah, G., & Eshet-Alkalai, Y. (2018). The contribution of text-highlighting to comprehension: A comparison of print and digital reading. *Journal of Educational Multimedia and Hypermedia*, *27*(2), 153–178.

Betancourt, M. (2017). A conceptual introduction to hamiltonian Monte Carlo. arXiv preprint arXiv:1701.02434.

Blanchard, J., & Mikkelson, V. (1987). Underlining performance outcomes in expository text. *The Journal of Educational Research*, *80*(4), 197–201.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language (Vol. 76) (No. 1). New York, NY: Columbia University Press.

Cashen, V. M., & Leicht, K. L. (1970). Role of the isolation effect in a formal educational setting. *Journal of Educational Psychology*, *61*(6, Pt.1), 484–486.

Collins, M., Dasgupta, S., & Schapire, R. E. (2002). A generalization of principal components analysis to the exponential family. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (pp. 617–624). Cambridge, MA: MIT Press. Retrieved from http://papers.nips.cc/paper/2078-a-generalization-of-principal-components-analysis-to-the-exponential-family.pdf

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

Crouse, J. H., & Idstein, P. (1972). Effects of encoding cues on prose learning. *Journal of Educational Psychology*, *63*(4), 309.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58.

Faw, H. W., & Waller, T. G. (1976). Mathemagenic behaviours and efficiency in learning from prose materials: Review, critique and recommendations. *Review of Educational Research*, *46* (4), 691–720.

Fowler, R. L., & Barker, A. S. (1974). Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology*, *59*(3), 358.

Gantner, Z., Drumond, L., Freudenthaler, C., Rendle, S., & Schmidt-Thieme, L. (2010). Learning attribute-to-feature mappings for cold-start recommendations. In 2010 IEEE International Conference on Data Mining (pp. 176-185). IEEE

Harp, S. F., & Mayer, R. E. (1997). The role of interest in learning from scientific text and illustrations: On the distinction between emotional interest and cognitive interest. *Journal of Educational Psychology*, *89* (1), 92.

Hartley, J., Bartlett, S., & Branthwaite, A. (1980). Underlining can make a difference—Sometimes. *The Journal of Educational Research*, *73*(4), 218–224.

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

Idstein, P., & Jenkins, J. R. (1972). Underlining versus repetitive reading. *The Journal of Educational Research*, *65* (7), 321–323.

Johnson, L. L. (1988). Effects of underlining textbook sentences on passage and sentence retention. *Literacy Research and Instruction*, *28*(1), 18–32.

Kintsch, W. (1980). Learning from text, levels of comprehension, or: Why anyone would read a story anyway. *Poetics*, *9*(1–3), 87–98.

Klare, G. R., Mabry, J. E., & Gustafson, L. M. (1955). The relationship of patterning (underlining) to immediate retention and to acceptability of technical material. *Journal of Applied Psychology*, *39*(1), 40.

Koren, Y., Bell, R., & Volinsky, C. (2009). *Matrix factorization techniques for recommender systems*, Washington, DC: IEEE.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14* (2), 219–224.

Lan, A. S., Waters, A. E., Studer, C., & Baraniuk, R. G. (2014). Sparse factor analysis for learning and content analytics. *The Journal of Machine Learning Research*, *15*(1), 1959–2008.

Landgraf, A. J., & Lee, Y. (2015). Dimensionality reduction for binary data through the projection of natural parameters. arXiv preprint arXiv:1510.06112.

Lee, S., Huang, J. Z., & Hu, J. (2010). Sparse logistic principal components analysis for binary data. *The Annals of Applied Statistics*, *4*(3), 1579.

Leutner, D., Leopold, C., & Elzen-Rump, D. (2007). Self-regulated learning with a text-highlighting strategy: A training experiment. *Zeitschrift Fu¨r Psychologie/Journal of Psychology*, *215*(3), 174.

Lonka, K., Lindblom-YlÄnne, S., & Maury, S. (1994). The effect of study strategies on learning from text. *Learning and Instruction*, *4*(3), 253–271.

Lorch, E. P., & Klusewitz, M. A. (1995). Effects of typographical cues on reading and recall of text. *Contemporary Educational Psychology*, *20*(1), 51–64.

Lutz, J., Briggs, A., & Cain, K. (2003). An examination of the value of the generation effect for learning new material. *The Journal of General Psychology*, *130*(2), 171–188.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (pp. 55-60).

Masson, M. E., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, *57*(3), 203.

McNamara, D. S. (1995). Effects of prior knowledge on the generation advantage: Calculators versus calculation to learn simple multiplication. *Journal of Educational Psychology*, *87*(2), 307.

Menon, A. K., & Elkan, C. (2011). Link prediction via matrix factorization. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.), *Machine learning and knowledge discovery in databases* (pp. 437–452). Berlin: Springer.

Mills, C., Graesser, A., Risko, E. F., & D'Mello, S. K. (2017). Cognitive coupling during reading. *Journal of Experimental Psychology: General*, *146*(6), 872.

Miyatsu, T., Nguyen, K., & McDaniel, M. A. (2018). Five popular study strategies: Their pitfalls and optimal implementations. *Perspectives on Psychological Science*, *13*(3), 390–407.

Nist, S. L., & Hogrebe, M. C. (1987). The role of underlining and annotating in remembering textual information. *Literacy Research and Instruction*, *27*(1), 12–25.

Nist, S. L., & Kirby, K. (1989). The text marking patterns of college students. *Reading Psychology: An International Quarterly*, *10*(4), 321–338.

Pavlik, P. I., Cen, H., & Koedinger, K. R. (2009). *Performance factors analysis: A new alternative to knowledge tracing*. Amsterdam: IOS Press. Retrieved from http://dl.acm.org/citation.cfm?id=1659450.1659529

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

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543). Baltimore, MD: Association for Computational Linguistics.

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

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J.. (2015). Deep knowledge tracing. Curran Associates, Inc., Retrieved from http://papers.nips.cc/paper/5654-deep-knowledge-tracing.pdf

Rasch, G. (1980). Probablistic models for some intelligence and attainment tests.

Rickards, J. P., & August, G. J. (1975). Generative underlining strategies in prose recall. *Journal of Educational Psychology*, *67*(6), 860.

Rye, C., Wise, R., Jurukovski, V., Desaix, J., & Avissar, Y. (2016). *Biology*, Houston, TX: . OpenStax.

Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). *Methods and metrics for cold-start recommendations*. New York: ACM. https://doi.org/10.1145/564376.564421

Shi, Y., Larson, M., & Hanjalic, A. (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), 3.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592.

Wallace, W. P. (1965). Review of the historical, empirical, and theoretical status of the von Restorff phenomenon. *Psychological Bulletin*, *63*(6), 410.

Wollen, K. A., Cone, R. S., Britcher, J. C., & Mindemann, K. M. (1985). *The effect of instructional sets upon the apportionment of study time to individual lines of text*. (Vol. 4) (No. 2). .

Yue, C. L., Storm, B. C., Kornell, N., & Bjork, E. L. (2015). Highlighting and its relation to distributed study and students' metacognitive beliefs. *Educational Psychology Review*, *27*(1), 69–78.

Zhang, L., Xiong, X., Zhao, S., Botelho, A., & Heffernan, N. T. (2017). Incorporating rich features into deep knowledge tracing. In J. Reich, C. Thille, & C. Urrea (Eds.), *Proceedings of the fourth (2017) ACM conference on learning @ scale* (pp. 169–172). New York: ACM. https://doi.org/10.1145/3051457.3053976