# Exploring Document Retrieval Features Associated with Improved Short- and Long-term Vocabulary Learning Outcomes

Rohail Syed
School of Information
University of Michigan
Ann Arbor, Michigan
rmsyed@umich.edu

Kevyn Collins-Thompson
School of Information
University of Michigan
Ann Arbor, Michigan
kevynct@umich.edu

## ABSTRACT

A growing body of information retrieval research has studied the potential of search engines as effective, scalable platforms for self-directed learning. Towards this goal, we explore document representations for retrieval that include features associated with effective learning outcomes. While prior studies have investigated different retrieval models designed for teaching, this study is the first to investigate how document-level features are associated with actual learning outcomes when users get results from a personalized learning-oriented retrieval algorithm. We also conduct what is, to our knowledge, the first crowdsourced longitudinal study of *long-term* learning retention, in which we gave a subset of users who participated in an initial learning and assessment study a delayed post-test approximately nine months later. With this data, we were able to analyze how the three retrieval conditions in the original study were associated with changes in long-term vocabulary knowledge. We found that while users who read the documents in the personalized retrieval condition had immediate learning gains comparable to the other two conditions, they had better long-term retention of more difficult vocabulary.

**ACM Reference Format:**
Rohail Syed and Kevyn Collins-Thompson. 2018. Exploring Document Retrieval Features Associated with Improved Short- and Long-term Vocabulary Learning Outcomes. In *CHIIR '18: 2018 Conference on Human Information Interaction & Retrieval, March 11–15, 2018, New Brunswick, NJ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3176349.3176397

## 1 INTRODUCTION

Recent work in information retrieval has focused on models, algorithms, and evaluation methods at the intersection of general Web search with learning-oriented intents [6, 18] to investigate different dimensions of the concept of search as learning. For example, some studies have investigated and demonstrated the evident demand for using Web search engines for accomplishing learning or discovery goals [2, 7, 13]. Other studies have investigated the use of Web search engines to accomplish learning goals and possible links between search behavior and learning outcomes [1, 6, 9, 23]. Additional research also investigated effects of search behavior on learning outcomes but through more indirect measures of learning, such as classifying users as beginners or experts and observing their behaviors [10, 22].

Learning outcomes have also served as the specific basis for a retrieval objective or framework within educational applications. Prior work by Collins-Thompson and Callan on the REAP system [5] demonstrated how a system could automatically crawl, filter, curate, and retrieve a set of Web documents to accommodate learning a predefined set of words from context within an intelligent tutoring system. However, this framework did not extend to supporting a real-time search engine for finding personalized learning-oriented documents for arbitrary topics or ad-hoc queries. One of the first studies to introduce a retrieval model whose objective specifically aimed at optimizing learning outcomes was the study by Syed and Collins-Thompson [19], later extended in [20]. In that study, the authors ran a large-scale crowdsourced user study to investigate actual changes in knowledge states of participants before and after they were provided personalized documents to read [20]. They demonstrated that their personalized retrieval approach could achieve better learning gains per unit of effort compared to a commercial search engine baseline.

That work, however, did not explore which specific features in documents or document sets being retrieved were likely to help or hinder learning, and we know of little work in general on that question. Such features might include the number and density of accompanying images, the difficulty of the text, the length of paragraphs, and so on. In this work, we study an extensive set of features based on a dataset from the original study by Syed and Collins-Thompson to determine what features best predict different learning outcomes, as well as a few other important learning-related variables such as time spent reading. We also assess the long-term retention of those who took part in the original study by Syed and Collins-Thompson [20] by conducting a delayed post-test with a subset of those users.

The main contributions of this work are as follows: (1) We investigate a comprehensive set of document, document-set and user interaction features for their association with a variety of short-term and long-term learning measures on a vocabulary learning task; (2) using predictive models based on these features and outcomes, we show that even models without user-specific information are somewhat effective at predicting which documents are likely to be associated with improved learning, with user-specific features further improving model fit; (3) We conduct the first study of long-term retention in the context of Web search for learning; and (4) We investigate temporal changes in knowledge state from three stages of learning (pre-test, immediate post-test, delayed post-test) and connect these outcomes to properties of the original study conditions. Finally, we discuss the implications of our findings for further improving search-as-learning frameworks.


*CHIIR '18, March 11–15, 2018, New Brunswick, NJ, USA*

ACM ISBN 978-1-4503-4925-3/18/03…$15.00
https://doi.org/10.1145/3176349.3176397

## 2 RELATED WORK

A number of prior studies have investigated different aspects of how Web search can be, and is, used for learning [2, 7, 10, 14, 22, 24]. Several large-scale query log studies have investigated how search interaction features, including queries issued, pages visited, time spent and more, varied between experts and non-experts [10, 22] and between expert and non-expert website *content* [14]. Work by Kim et al. [14] also found evidence of users exhibiting "stretch reading" behavior where they choose to read documents significantly higher than their own expected understanding level. A recent study by Verma et al. [21] also found that readability of a document showed a negative correlation with how easily users could find relevant content on the page. This suggests that certain features of Web page content could have an effect on which pages a user chooses to visit when learning, and their ability to find relevant information on that page.

**Short-term learning.** While such studies are useful in understanding estimated changes in a user's domain expertise, they did not directly assess actual learning outcomes of participants engaging in a learning task. There have been quite a few studies that have investigated this [1, 6, 9, 12, 20, 24]. Several studies have examined how users in largely unconstrained search environments exhibit different search behaviors, which in turn may link to changes in assessed learning outcomes [9, 12, 24]. Other studies assessed changes in knowledge state as a function of search behavior and documents that were selected or preferred by the user, e.g. via exploring intrinsically diverse results [6], scanning multiple documents first, ordering and then reading them [1], or simply providing a custom set of documents without taking query input at all from the user [20]. Furthermore, very few studies that have explicitly assessed Web search and its intersection with learning outcomes have tried to directly optimize the utility of search results for learning [5, 20]. One of the primary goals of this study is to explore how document properties, used as features in robust regression models, are associated with actual learning outcomes, and predictive of future learning outcomes as part of optimal content retrieval by search-for-learning systems.

**Long-term learning.** Despite the importance of long-term retention as a learning goal, few, if any, studies to our knowledge have considered search engines aimed at long-term retention of knowledge, or long-term gains in knowledge. In practice, learning is a continuous process, constantly engaged as a function of information that we observe and cognitively process [4]. While the distinction between short and long-term learning has been established for quite some time [3], its application to a search retrieval framework would be a novel and critical contribution. Recent work by Eickhoff et al. [10] investigated how users' domain expertise changes over time in a large-scale query log analysis, but did not measure actual knowledge of the users at any stage. Earlier work by Wildemuth [23] assessed how Web search tactics and behaviors changed during a nine-month span in the context of an educational course. The author investigated different search patterns and differences in actual learning outcomes at three separate temporal stages. While this offered useful insight into how search behavior may change with changes in knowledge, it did not give insight as to what properties of the documents were influencing these changes, nor did it investigate different retrieval algorithms and their possible effect on the learning changes. In this study, we conduct the first crowdsourced longitudinal study to assess a participant's long-term retention of knowledge, based on measured learning outcomes from an earlier user study of vocabulary learning. We investigate what document and user variables may be associated with long-term changes in knowledge state and how different document retrieval algorithms may influence the strength of these changes.

## 3 DATASETS

To conduct our analysis, we used a dataset provided from earlier work by Syed and Collins-Thompson [20], who conducted a large-scale crowdsourced user study to evaluate the effectiveness of their retrieval models for personalized learning. They tested their approach on 10 topics, with 40 participants in each condition, yielding a total of 863 judgments after enforcing quality control filters. Participants first completed a pre-test consisting of 10 multiple-choice questions on definitions of related vocabulary keywords. They were then provided a personalized set of documents that they had to read and were finally given a post-test, identical to the pre-test.

In this study, we only consider participants who were shown different personalized sets of documents, as this allows us to compare changes in document features to changes in knowledge state. This reduced dataset contains 283 records, with each record containing data about a unique participant, including their prior and post knowledge scores for each of the 10 keywords, the time they spent in the reading section, and the set of documents they were provided (and read).

In addition to analysis on this dataset, we investigated the effect of a variety of features on robust, or long-term, learning outcomes by conducting a follow-up crowdsourced test. As we could not control which participants would return, especially since it had been nearly nine months since the original study, we did not get long-term data for all 863 participant records, though we were still able to gather a reasonable number of records. In the following section, we will first investigate the features that best predict different measures of learning outcomes along with time spent reading. We will then give an analysis of data for long-term learning outcomes and their relationship to document and user features.

## 4 ANALYSIS

Overall, we considered a set of document features that included features pertaining to image use, vocabulary difficulty, word count and content structure (described in Section 4.1). A complete list, including user-dependent features, can be found in Table 1. In this section we analyze the relationships between these features and a variety of measures of learning (Section 4.2). We fit and analyze models restricted to user-independent features (Section 4.3) and then with all features (Section 4.4).

### 4.1 Choice of Features

We chose document and user features based on various concepts investigated in earlier studies [8, 11, 16, 20, 21, 25]. Broadly, the features we chose can be grouped as follows:

1. **Image content.** Some studies have found that providing plain-text filtered documents (with images removed) improves learning outcomes [11] over the original document, possibly suggesting a negative effect of image use in Web documents on learning. However, other studies found positive

| Type | Group | Feature | Description |
|---|---|---|---|
| D | Effort | $WordCount$ | Total number of unigrams in the document. |
| D | Effort | $KeyCount$ | Total number of keywords in the document. |
| D | Effort | $DocumentCount$ | Total number of documents in the set. This feature ranges from 1 to 10. |
| D | Effort | $WordsPerDocument$ | Ratio of $WordCount$ to $DocumentCount$. |
| D | Effort | $DocumentAgeDifficulty$ | $85^{th}$ percentile Age-of-Acquisition score for the document. Uses the expanded set of scores from the study by Kuperman et al. [15]. |
| D | Effort | $WeightedWordCount$ | Each unigram is assigned its corresponding "age" from the Age-of-Acquisition dataset. These scores, for each occurrence of each unigram in the document, are summed. |
| D | Effort | $AverageParaLength$ | Average length of each paragraph in the document. Computed as count of all unigrams in all HTML <p> tags divided by total instances of <p> tags. |
| D | Images | $ImageCountTag$ | Total instances of the HTML <img> tag that appeared in the document. |
| D | Images | $ImageCountManual$ | Total instances of non-advertising and non-navigational images that appeared in the document. Counted manually. |
| D | Images | $ImageToText$ | Ratio of $ImageCountTag$ to $WordCount$. |
| D | Links | $OutboundLinks$ | The count of all outbound links. |
| D | Keywords | $KeywordDensity$ | Computed as the count of occurrences of any of the $N$ keywords $k_1, \ldots, k_N$ divided by the count of all words (i.e. $WordCount$). |
| D | Keywords | $WeightedDensity$ | Same as $KeywordDensity$ except the denominator is the $WeightedWordCount$ feature. |
| U+D | Keywords | $IncorrectKeysRatio$ | Total occurrences of keywords that the participant got wrong in their pre-test, divided by the total occurrences of any keyword in that document. |
| U+D | Keywords | $IncorrectSemanticRatio$ | First compute $SRel$ scores: the relevance of each keyword instance in a document computed as the average Word2Vec similarity [17] of its five surrounding words (both ahead and behind). $IncorrectSemanticRatio$ is the sum of all $SRel$ scores for keywords the participant got wrong on the pre-test, divided by the total sum of $SRel$ scores. |
| DS | Keywords | $LogWeightedDensity$ | Same as $WeightedDensity$ except that instead of simply summing the values over the set of documents, each successive document's value of $WeightedDensity$ was reduced by a DCG discount factor of $\log_2(p+1)$ where $p$ is the rank in the set of documents. |
| DS | Images | $Set\_ImageToText$ | Set-level calculation of $ImageToText$. |
| DS | Effort | $Set\_AvgParaLength$ | Set-level calculation of $AverageParaLength$. |
| DS | Keywords | $Set\_KeyDensity$ | Set-level calculation of $KeywordDensity$. |
| DS | Keywords | $Set\_WeightDensity$ | Set-level calculation of $WeightDensity$. |
| U+DS | Keywords | $Set\_IncorrectRatio$ | Set-level calculation of $IncorrectKeysRatio$. |
| U+DS | Keywords | $Set\_IncorrectSemsRatio$ | Set-level calculation of $IncorrectSemanticRatio$. |
| U+DS | Keywords | $ExpectedKnowledge$ | Expected knowledge computed as a personalized sigmoid function of keywords [20]. |
| U | | $PriorKnowledge$ | Sum of initial correct answers to the vocabulary terms needed to be learned. |

**Table 1: Set of features that were considered. "U" are User features that involve prior data about the User's knowledge. "D" are Document features that require only individual document data. "DS" are Document Set features based on treating the set of documents as a single bag-of-words. The "D" features values were aggregated by summation, since learning outcomes were measured against sets of documents.**

association of image use and learning, when used appropriately [16] and a positive association with the fraction of images in documents and the ability of users to find relevant content [21].

2. **Keyword content.** Prior work has found that optimizing document selection by difficulty-weighted keyword density improved multiple measures of learning outcomes [20] in a vocabulary learning task where the system determined the set of keywords that a participant had to learn. We also investigate other keyword features like the count of occurrences of keywords unknown to the user relative to all keywords.
3. **Effort.** Prior work has suggested that too much effort on the part of users can be overwhelming and, according to Cognitive Load Theory, could hurt learning outcomes [8]. On the other hand, having "desirable difficulties" [3] has been found to improve learning outcomes. We consider effort as functions of document count, word count and reading-difficulty-weighted measures of content.
4. **Embedded links.** Several studies have found that embedded links in documents can disturb the linearity of the learning process [25] and can add extra cognitive load [8], potentially hurting learning gains.

## 4.2 Measures of Learning Outcomes

We now evaluate the following measures of learning outcomes, on the provided sets of $K = 10$ vocabulary questions, with $Pre_k$ as

prior knowledge of keyword $k$, $Post_k$ as corresponding post knowledge and $r_k$ as vocabulary difficulty level of $k$:

**Learning Gains (LG).** As a simple measure of learning growth we compute the total instances where a participant did not know a keyword to be learned in the pre-reading test and did know the definition in the post-reading test.

$$LG = \sum_{k=1}^{K} \begin{Bmatrix} 1 & Pre_k = 0 \text{ and } Post_k{=}1 \\ 0 & \text{otherwise} \end{Bmatrix}$$

**Difficulty-Weighted Gains (DWG).** This meausure is essentially the same as Learning Gains but we weight the learning gains of each keyword by the vocabulary difficulty level associated with it. These difficulty scores are retrieved from the expanded dataset from work by Kuperman et al. [15]. By weighting the learning gains by vocabulary difficulty, we can capture the intuition that learning more difficult words like 'luciferase' and 'eclogite' may require different features than those required for learning easier words like 'minerals' or 'soils'.

$$DWG = \sum_{k=1}^{K} r_k \begin{Bmatrix} 1 & Pre_k = 0 \text{ and } Post_k{=}1 \\ 0 & \text{otherwise} \end{Bmatrix}$$

**Realized Potential Gains (PG).** This is a measure of the participant's actual Learning Gain relative to their maximum possible Learning Gain. Specifically, for a set of 10 vocabulary terms being tested, we have:

$$PG = \frac{LG}{10 - \sum_{k=1}^{10} Pre_k}$$

Participants who had perfect prior knowledge (10/10) were omitted from analysis as they could not have theoretically shown any improvement.

**Final Knowledge (FK).** This is a much simpler measure of learning outcome where we take the linear sum of the participant's final test scores, regardless of their prior performance. Specifically, we have:

$$FK = \sum_{k=1}^{K} Post_k$$

**Learning Hindrance (LH).** While previous measures of learning outcomes assessed positive learning outcomes, it is also important to understand features that may *hinder* learning. We consider Learning Hindrance to be the total keywords that a participant got wrong in the pre-test and got wrong again on the post-test, indicating that they were unable to learn the definition. Specifically, we have:

$$LH = \sum_{k=1}^{K} \begin{Bmatrix} 1 & Pre_k = 0 \text{ and } Post_k{=}0 \\ 0 & \text{otherwise} \end{Bmatrix}$$

**Total Reading Time (TR).** While this is not technically a measure of learning, it is an important measure to analyze as it can help determine what document and user features influence how much or how little time people are willing to spend when engaged in a learning task. This is measured as the total time (ms) a user spent reading the set of documents they were provided.

## 4.3 Prediction without User Data

There are many scenarios in Web search where it may be difficult or impossible to obtain an accurate assessment of a user's prior knowledge, especially for any arbitrary topic. Thus, here we investigate document features that are completely independent of the user ("D" and "DS" type properties only) and assess how well robust regression models trained on these features can predict learning outcomes. These models could facilitate learning-oriented retrieval for situations where a Web search framework has access to document data but not to a user's prior knowledge.

In selecting the features for each model, we applied a stepwise algorithm using AIC (Akaike information criterion) to reduce the likelihood of overfitting to unnecessary features. We used min-max scaling to normalize the predictor and dependent variables in all models. To reduce the effect of any specific influential points on model fitting, we fit all the models with robust regression. We tabulate the trained models and averaged 10-fold cross-validated correlations in Table 2.

The results from Table 2 show that even without any features about the user, we can still get reasonably strong correlations between predicted learning outcomes and actual outcomes. For learning gains, the Difficulty-Weighted Gains tend to show substantially better improvement over the unweighted gains. On the other hand, the Final Knowledge state variable shows a much stronger correlation as does the Learning Hindrance variable. We visualize the trained models for Difficulty-Weighted Gains, Final Knowledge and Learning Hindrance in Figure 1.

For the selected features, all positive measures of learning showed positive weights for ImageCountManual and negative weights for ImageCountTag, suggesting that, in general, Web pages having more relevant images tend to be associated with better actual learning outcomes with the opposite being true for irrelevant images (such as ads and navigational icons), possibly due to their distracting to the user. This is consistent with existing work in this area that has suggested that having images in learning material has been found to both help and harm learning outcomes, depending on the study[8]. All measures of learning gains showed a negative relationship with the total number of links in the document, which is consistent with what we would have expected from theory (Section 4.1). However, it is not entirely clear why Final Knowledge shows a positive relationship with total links. We also observe that both unweighted and weighted learning gains measures were positively affected by weighted keyword density, at the individual document level. This is consistent with the results from [20] that found that document sets produced by greedy document-level optimization for weighted keyword density outperformed commercial baseline results in terms of learning gains. However, we also found that at the set level, the weights for weighted density were negative. This disparity may be due to the document-level features being computed as sums across all documents in the set, thus making the DocumentCount feature an implicit feature in document-level weighted density. This suggests that at the set level, keyword density should be rewarded but weighted keyword density should be penalized. The positive learning outcomes from the earlier study [20] could be attributed to the fact that set-level weighted density was also strongly negatively correlated to features like ImageCountTag and OutboundLinks, so that optimizing towards weighted density could have indirectly brought out higher quality documents.

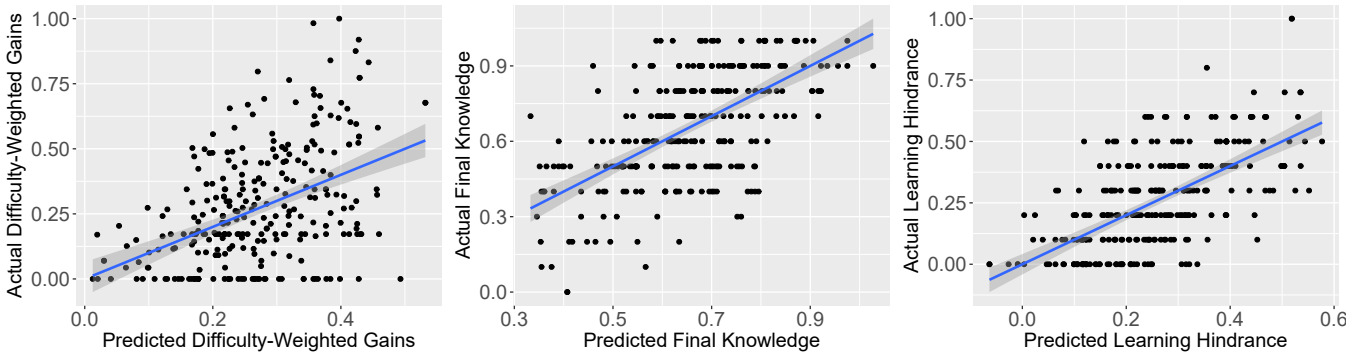


**Figure 1: Predicted and actual learning measures trained on non-user features.**

| Feature | LG | DWG | PG | FK | LH | TR |
|---|---|---|---|---|---|---|
| WordCount | | 0.4379 | 3.6121 | | -0.5535 | -2.8926 |
| WeightedWordCount | | | -3.5873 | | | 2.3241 |
| AverageParaLength | | | -0.2336 | -0.2755 | 0.3486 | |
| ImageCountManual | 0.2904 | 0.3224 | 0.5441 | 0.2996 | -0.2738 | 0.1544 |
| OutboundLinks | -0.2394 | -0.3990 | | 0.2681 | -0.1498 | 0.3157 |
| KeywordDensity | -2.4830 | -1.9237 | -1.9809 | | | |
| WeightedDensity | 1.7599 | 1.8847 | 2.1748 | | | |
| DocumentAgeDifficulty | | 0.3747 | -0.2834 | -0.2651 | 0.3308 | |
| ImageToText | | | | | | |
| ImageCountTag | -0.3068 | -0.2283 | -0.6259 | -0.2909 | 0.2688 | 0.1498 |
| KeyCount | | -0.4071 | | | | |
| LogWeightedDensity | 0.5371 | | | | | |
| DocumentCount | 0.4221 | | | | | 0.0864 |
| WordsPerDoc | | | | | 0.4481 | |
| Set_AvgParaLength | 0.1492 | 0.1832 | 0.2393 | 0.1142 | -0.1181 | 0.2591 |
| Set_ImageToText | | | | -0.2189 | 0.1909 | -0.2600 |
| Set_KeyDensity | 1.5808 | 2.0079 | 1.6829 | -0.3255 | 0.2626 | |
| Set_WeightDensity | -1.5624 | -1.8801 | -2.1677 | | | |
| **Performance** | **LG** | **DWG** | **PG** | **FK** | **LH** | **TR** |
| Correlation (model prediction vs actual) | 0.3296 | 0.3611 | 0.3436 | 0.5810 | 0.6117 | 0.2376 |

**Table 2: Trained normalized features for different dependent variables. Values for corresponding features are learned coefficients in the robust regression model. LG = Learning Gains; DWG = Difficulty-Weighted Gains; PG = Potential Gains; FK = Final Knowledge; LH = Learning Hindrance; TR = Total Reading Time (ms).**

We find a similar tradeoff when it comes to average paragraph length, with the set-level average (micro-averaged across documents) being positively correlated with all measures of learning, suggesting less segmentation of the text can be an indication of higher-quality content for learning. However, we also note that the document-level average showed the opposite trend for Potential Gains and Final Knowledge, possibly suggesting that average paragraph lengths should be longer but there should be fewer documents overall.

Finally, we note that the models are not simply capturing the intuition that having more documents results in stronger gains. The "DS" set-level features are mostly ratio features which are invariant to proportional increases in the amount of content but are dependent on the *relative* changes of different types of content. Adding or removing documents to a set would give no guarantee of increasing or decreasing these values (e.g. Set_AvgParaLength had nearly 0 correlation with DocumentCount). If we excluded all the "DS" features, the new correlated strength decreased by about 16.3% averaged across the six models and if we consider *only* DocumentCount as a feature, the drop is substantially higher at 38.9%, suggesting that the models explain more than just exposure to more content is associated with higher learning outcomes.

## 4.4 Predicting with User Data

We have seen that in the absence of user-dependent features, we were able to train robust regression models on multiple measures

| Feature | LG | DWG | PG | FK | LH | TR |
|---|---|---|---|---|---|---|
| WordCount | | | | | | -2.5116 |
| WeightedWordCount | | | | | | 1.8478 |
| AverageParaLength | -0.1523 | | | | 0.1066 | |
| ImageCountManual | 0.3077 | 0.3867 | 0.5178 | 0.2353 | -0.2154 | |
| OutboundLinks | | | | | | |
| IncorrectSemanticRatio | | | 0.7476 | | | 0.6536 |
| KeywordDensity | -0.4643 | -0.5915 | -2.2101 | -0.5441 | 0.3250 | -0.2334 |
| WeightedDensity | | | 2.2856 | | | |
| DocumentAgeDifficulty | | | -0.4410 | | | |
| ImageToText | | | | | | |
| IncorrectKeyRatio | 0.3443 | 0.3578 | | 0.3933 | -0.2410 | -0.5565 |
| ImageCountTag | -0.1759 | -0.2824 | -0.5097 | -0.1426 | 0.1231 | 0.2191 |
| KeyCount | | | | | | 0.3497 |
| LogWeightedDensity | 0.3261 | 0.4702 | | 0.3570 | -0.2283 | |
| DocumentCount | | | | | | 0.2834 |
| WordsPerDoc | | | | | | |
| ExpectedKnowledge | -0.1199 | | | -0.1757 | 0.0839 | -0.2341 |
| Set_AvgParaLength | 0.1466 | | | 0.1098 | -0.1026 | 0.2404 |
| Set_ImageToText | -0.2745 | -0.1738 | -0.3182 | -0.2347 | 0.1921 | -0.1901 |
| Set_KeyDensity | | 1.4125 | 1.3909 | | | |
| Set_WeightDensity | | -1.4657 | -1.7781 | | | |
| Set_IncorrectRatio | -0.6198 | -0.2546 | -0.4914 | -0.6803 | 0.4338 | |
| Set_IncorrectSemsRatio | 0.4063 | | | 0.4612 | -0.2844 | |
| PriorKnowledge | -0.3694 | -0.3889 | 0.3289 | 0.7584 | -0.6414 | 0.3565 |
| **Performance** | **LG** | **DWG** | **PG** | **FK** | **LH** | **TR** |
| Correlation (model prediction vs actual) | 0.4571 | 0.5091 | 0.3908 | 0.7156 | 0.7499 | 0.2650 |
| Robust correlation with PriorKnowledge | 0.3744 | 0.3397 | 0.2731 | 0.6657 | 0.7361 | -0.0563 |

**Table 3: Trained normalized features for different dependent variables (considering *all* possible features). Values for corresponding features are learned coefficients in the robust regression model.**

of learning, resulting in observed trends that were commensurate with findings from existing literature. Now we attempt to further augment the power of these results by including all the features from Table 1 in our model. Repeating the same feature selection and model fitting process as before, we have the results in Table 3.

We first note that including all features improved the cross-validated correlations for all measures of learning, and for some quite substantially. This is not unexpected, given that we are adding signals which have a naturally strong correlation to most measures of learning already. For example, regardless of other properties, the user's prior knowledge could be expected to have a strong negative correlation with Learning Gains since users with higher prior knowledge naturally have less opportunities for improvement. Indeed, we trained the set of six learning measures against a robust model containing *only* PriorKnowledge as a predictor and found substantially strong correlations from that alone (last row of Table 3). However, training against the full set of features did show significant improvement in predicting Learning Gains, Difficulty-Weighted Gains, Potential Gains and especially Total Reading, which had almost no correlation with PriorKnowledge. As such, there are definitely advantages to incorporating both user features and document features for better results.

Using all features, we see similar trends to those we saw before: (1) all measures of learning outcomes had positive coefficients for the count of relevant images, and those measures that had count of all images as a significant feature had negative weights; (2) weighted keyword density again shows conflicting association with learning outcomes at the set level vs. the sum of document level; (3) we see a similar effect that we discussed earlier with average paragraph lengths as well as with total embedded links.

However, we also notice some new effects and features. First, the ImageToText ratio feature was in the original models, but was not significant for most of the features. In this set of all features, the set-level ImageToText feature has significant negative weight for *all* measures of learning, suggesting that in general, while more images might be helpful, there needs to be an overall balance between how many images there are per unit of text. Second, the ratio of counts of unknown keywords to all keywords is a positive predictor of better learning outcomes at the document level. However, it shows the opposite trend at the set level, either suggesting that in aggregate a set of documents should *not* have stronger coverage of unknown

keywords (that need to be learned). The reasons for this require further study.

In aggregate, this enhanced set of features has given us trained models that do show expected improvements over the document-features-only models and much of the same observations remain valid in these new models as well. While Syed and Collins-Thompson [20] demonstrated strong improvements in learning efficiency (learning gains per unit of effort), the models introduced here may lead to improvements in learning effectiveness (learning gains or final knowledge state) or strong reductions in learning hindrance.

## 5 LONG-TERM RETENTION

We now describe a crowdsourced longitudinal study of long-term retention, or *robust learning*, in which a subset of users who participated in an initial learning and assessment study also completed a delayed post-test nine months later, in order to study how much of their original word learning they had retained over time.

### 5.1 Study design

Our experiment used the same platform, Crowdflower, as the study by Syed and Collins-Thompson [20], as well as the original crowd response dataset from that study. Our study design included three pages of multiple-choice question tests for three topics out of the ten total that were originally tested. Afterwards, participants completed a Likert-scale survey of the perceived importance of various "learning factors" [1] on learning outcomes.

We limited this delayed post-reading assessment to only three topics to prevent participants from having to take too many tests and possibly having fatigue influence the results. We retained explicit quality control measures by adding gold standard test questions in each of the three tests that participants had to pass and we randomized the order in which the assessments appeared. Unfortunately, while the Crowdflower platform allows us to see the unique worker's ids after an experiment has terminated, they do not allow us to have this information during the experiment, nor do they allow us to specifically request certain workers. As such, we had to rely on chance that we would get repeat participants and further on chance that some of those repeat participants would have participated in one of the three selected topics. To maximize the number of data points we could get, we chose the three topics that had the lowest number of unique participants[1].

In the original study [20], we gathered a total of 1200 data points (judgments). In this study, we accumulated a total of 600 judgments from the crowd, within which we found 36 unique repeat participants who had taken part in the original study (out of a maximum of 116 from the set of three topics we chose) and there were 83 unique (participant, topic) tuples that matched the original dataset. After filtering out those who did not answer all the gold standard test questions correctly, we ended up with 81 unique tuples. We performed the subsequent analysis on this dataset, matched against the original dataset. For notation purposes, we consider "pre-test" to be the pre-reading test results from the original study, "post-test" to be the post-reading test results from the original study and "delayed-test" to be the test results from the (later) crowdsourced study described here.

[1]This increased the likelihood of getting more complete sets of (participant, topic) tuples across all topics.

| Difficulty Split | Lower Difficulty | Higher Difficulty | p-val |
|---|---|---|---|
| Robust Gains (Long-term) | 1.025 | 1.000 | 0.867 |
| Retained Gains | 0.457 | 0.765 | 0.002 |
| Retained Knowledge | 2.395 | 2.296 | 0.733 |
| Net Retained Knowledge | 1.815 | 1.160 | 0.067 |
| Learning Prior | 2.753 | 2.469 | 0.093 |
| Learning Gains (Short-term) | 0.679 | 1.296 | <.001 |

**Table 4: Averages for the two splits for each robust measure along with two short-term measures indicates better opportunity for gains in difficult terms.**

| Measure | Web | NP | P | p-val |
|---|---|---|---|---|
| Robust Gains | 1.960 | 2.000 | 2.136 | 0.809 |
| Retained Gains | 1.280 | 1.059 | 1.409 | 0.856 |
| Retained Knowledge | 4.440 | 4.706 | 4.955 | 0.706 |
| Net Retained Knowledge | 2.520 | 2.941 | 3.545 | 0.439 |
| Post-Test | 6.360 | 6.471 | 6.364 | 0.966 |
| Delayed-Test | 5.560 | 6.118 | 6.091 | 0.764 |

**Table 5: Averages of short- and long-term knowledge state measures, broken down by retrieval models.**

### 5.2 Robust learning outcomes

We consider the following measures of robust, or long-term, learning outcomes: (1) robust learning gains; (2) robust retention of learning gains; (3) robust retention of post-test knowledge and (4) robust change in post-test knowledge[2]. We define these measures as follows:

(1) **Robust Learning Gains.** Computed as the sum of keywords that a participant did not know in the pre-test and did know in the delayed-test.
(2) **Retained Gains.** Computed as the sum of keywords a participant learned (as defined by Learning Gains in Section 4) and that they still knew in the delayed-test.
(3) **Retained Knowledge.** Computed as the sum of keywords that a participant did get correct in the post-test and still got correct in the delayed-test.
(4) **Net Retained Knowledge.** Computed as signed sum of retentions in post-test knowledge (retention is positive if participant got the keyword correct in post-test and again in delayed-test; retention is negative if participant got the keyword correct in post-test and wrong in delayed-test).

*5.2.1 Variation by Keyword Difficulty.* We first analyze how the average robust measures compare when considering the averages of the lowest-difficulty keywords only versus the averages of the highest-difficulty keywords only. We split the set of ten keywords into sets of five by a median split on their Age-of-Acquisition scores [15]. We then compute each of the robust measures as well as the pre-test scores on each of the sets and perform a Kruskal-Wallis test to test for significance. The results are presented in Table 4.

[2]In this section, *robust* learning refers to participant learning that is retained over the long term, not to be confused with the robust regression estimation method used in our predictive models.

We find that of the four robust measures, Retained Gains and Net Retained Knowledge showed significant differences in means: (lower mean = 0.457, higher mean = 0.765, p=.002) and (lower mean = 1.815, higher mean = 1.160, p=.067)[3] respectively. This suggests that in general, of the keywords participants were able to learn and remember, more of these were likely to be difficult ones. On the other hand, the opposite trend with Net Retained Knowledge suggests that overall participants were also more likely to *forget* the meanings of more difficult keywords. This shows an interesting balance where participants who retained short-term learning gains tended to retain acquired knowledge of more difficult terms better. However, in cases where they forgot newly-learned terms, they tended to lose acquired knowledge more with difficult terms as well. In aggregate, there appears to be more forgetting than retaining with difficult terms, suggesting that participants with better post-test knowledge of easier terms will likely show a better net retention of that knowledge even after a considerable time delay.

Another interesting finding is that the Robust Gains split was unaffected by difficulty but the short-term learning gains were strongly improved by higher difficulty (almost twice as much). We also observe that the averages of these measures suggest a negative relationship (i.e. lower short-term gains in easier terms led to better long-term gains of easier terms and vice versa for difficult terms). This may be explained by the fact that more difficult keywords are likely those that are more unfamiliar and novel to the learner and this novelty may facilitate better immediate recall but not long-term retention. Conversely, learning unknown but easier keywords may be less likely to cause learning gains as just a function of recall.

Related to the concept of *desirable difficulties* [3], it is possible that the easier keywords that were unknown to the participant were those that were difficult to learn but not so much that they inhibited long-term retention. This is supported by the Net Retained Knowledge results, where easier keywords showed substantially better net change in delayed-test knowledge. These results suggests that in personalizing document selection, it is important to incorporate the difficulty of unknown words.

*5.2.2 Variation by Retrieval algorithm.* We now analyze whether there were differences in robust learning outcomes depending on the search model a user was assigned in the original study. There were three possible models: (1) commercial search engine (**Web**); (2) non-personalized retrieval (**NP**) and (3) personalized retrieval (**P**). In our long-term dataset, each condition had roughly similar, but small, sample sizes (n=25, n=34, n=22) respectively. The **NP** and **P** algorithms exclusively considered a measure of difficulty-weighted keyword density as the document selection criteria, with **P** also incorporating information about the participants' prior knowledge and **NP** assuming zero prior knowledge for all participants. Details on these algorithms are provided in the original study [20].

We found that omnibus Kruskal-Wallis tests between these three models showed no significant differences for each of the four robust measures (Table 5), suggesting that in aggregate the choice of retrieval model didn't have significant impact on robust learning outcomes. However, if we split these features again by difficulty, we find some significant differences. In particular, both Robust Gains and Retained Gains showed significant differences (p<.05) when comparing only **Web** and **P** on higher difficulty keywords (Figure 3). In both cases, **P** outperformed **Web** (by 85% and 92% respectively),

[3]This significance was strengthened to p<.05 when normalizing by post-test knowledge

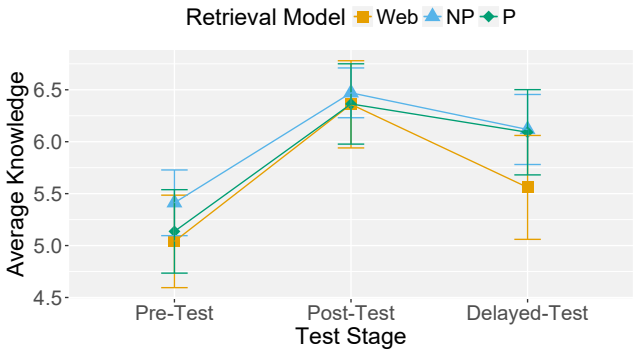


**Figure 2: Average changes in knowledge state (number of keywords correct) over three periods of assessment for each retrieval model.**

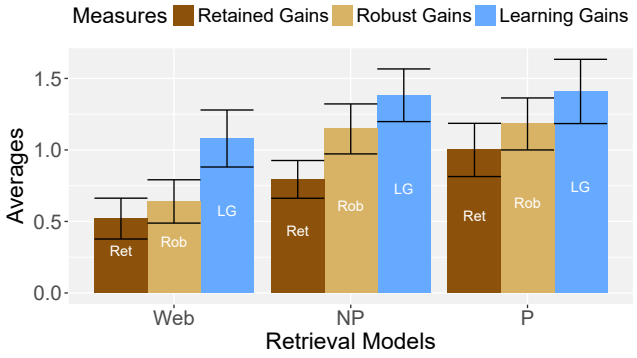


**Figure 3: For higher-difficulty keywords, the Personalized model (P) led to significantly better long-term retention of learned keywords than the baseline Web model.**

suggesting that the personalized algorithm introduced in [20] produced significantly better long-term improvements in knowledge of more difficult terms, including better retention of short-term gains on such terms.

We also observe some interesting variations in measures of final knowledge state. In particular, observe in Table 5 that the post-test final knowledge state showed very small differences across each of the models, suggesting that regardless of the retrieval model, the final knowledge state mostly ended up the same. However, in the delayed-test knowledge state, while there was consistent evidence of forgetting, this effect was distinctly stronger in **Web**, which was the commercial search baseline (Figure 2). This suggests that the other two models, proposed in [20] actually did demonstrate not just evidence of short-term improvements but very possibly evidence of long-term improvement as well.

Overall, we find that the personalized document retrieval model (Model **P**) showed substantially better ability compared to a commercial Web search model (Model **Web**) to help participants achieve long-term understanding of more difficult keywords and retain short-term learning gains of such keywords as well. We further find that, though not significant, the commercial model produced relatively stronger overall forgetting from post-test to delayed-test.

| Feature | Robust Gains | | | Net Retained Knowledge | | | Retained Gains | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Lower** | **Upper** | **Corr** | **Lower** | **Upper** | **Corr** | **Lower** | **Upper** | **Corr** |
| ImageCountManual | 2.912e+01 | 2.919e+01 | -.0875 | **3.265e+01** | **2.623e+01*** | **-.1901·** | 2.811e+01 | 2.972e+01 | -.0595 |
| OutboundLinks | 1.016e+03 | 9.165e+02 | -.0699 | **7.299e+02** | **1.15e+03*** | **.2777*** | 9.234e+02 | 9.765e+02 | .1015 |
| IncorrectSemanticRatio | **3.138e+00** | **4.614e+00†** | **.3930!** | **4.545e+00** | **3.531e+00·** | **-.3292†** | **3.21e+00** | **4.409e+00*** | **.4114!** |
| KeywordDensity | 4.044e-01 | 4.128e-01 | .0923 | **4.455e-01** | **3.788e-01*** | -.1363 | 3.832e-01 | 4.23e-01 | .0257 |
| WeightedDensity | 5.67e-02 | 5.801e-02 | .0949 | **6.3e-02** | **5.28e-02*** | -.1506 | 5.375e-02 | 5.942e-02 | .0262 |
| IncorrectKeyRatio | **3.214e+00** | **4.723e+00†** | **.3937!** | **4.632e+00** | **3.633e+00·** | **-.3154†** | **3.288e+00** | **4.513e+00*** | **.4032!** |
| KeyCount | 5.541e+02 | 5.191e+02 | -.1047 | **5.81e+02** | **4.941e+02·** | **-.2731*** | 5.32e+02 | 5.347e+02 | .0840 |
| LogWeightedDensity | 2.993e-02 | 3.065e-02 | .0870 | **3.353e-02** | **2.766e-02†** | **-.1983·** | 2.858e-02 | 3.128e-02 | -.0079 |
| ExpectedKnowledge | **9.48e+00** | **9.276e+00*** | **-.1915·** | 9.4e+00 | 9.329e+00 | **-.1952·** | **9.481e+00** | **9.298e+00·** | .0093 |
| Set_KeyDensity | 3.945e-02 | 4.104e-02 | .0310 | **4.647e-02** | **3.525e-02†** | **-.3245†** | 3.89e-02 | 4.116e-02 | -.0631 |
| Set_WeightDensity | 5.456e-03 | 5.667e-03 | .0286 | **6.494e-03** | **4.809e-03†** | **-.3195†** | 5.409e-03 | 5.668e-03 | -.0696 |
| Set_IncorrectRatio | **3.445e-01** | **5.035e-01*** | **.3200†** | **5.169e-01** | **3.694e-01*** | **-.4056!** | 3.687e-01 | 4.727e-01 | **.3151†** |
| Set_IncorrectSemsRatio | **3.375e-01** | **4.974e-01*** | **.3324†** | **5.112e-01** | **3.623e-01*** | **-.4074!** | 3.667e-01 | 4.639e-01 | **.3079†** |
| PriorKnowledge | **6.147e+00** | **4.553e+00!** | **-.5270!** | **4.378e+00** | **5.932e+00!** | **.4689!** | **5.964e+00** | **4.83e+00*** | **-.4452!** |
| **Survey Features** | | | | | | | | | |
| Novelty | **3.618e+00** | **4e+00*** | .0421 | **4.081e+00** | **3.636e+00*** | -.0797 | **3.571e+00** | **3.981e+00·** | .0494 |
| | Signif. codes: 0 '!' 0.001 '†' 0.01 '*' 0.05 '·' 0.1 ' ' 1 | | | | | | | | |

**Table 6: For each dependent variable (DV), "Lower" and "Upper" columns contain mean values for different features when considering either the subset of less than median of the DV or above median respectively. The third column "Corr" is the Pearson's correlation between each feature and each DV. Bold values are significant features at particular significance levels.**

*5.2.3 Analysis of median split.* In this section, we consider how each measure of robust learning differs, on average, with each of the features from Table 1 when considering two subsets of data, split on the median value of the corresponding robust measure. For space reasons, we only included features that showed significant differences or significant correlation. We tabulate the results in Table 6 and include the averages, the overall correlation of the feature with the measure, and the associated significance levels of both the splits and the correlations.

The first observation we make from these findings is that both Robust gains and Retained gains exclusively only showed significant differences on measures pertaining to the user's prior knowledge. This suggests that while short-term learning gains may be influenced by user-independent document features, neither long-term gains nor retention of short-term gains seem to be affected similarly. However, unlike what we saw in Section 4, here we note that both the set-level and sum of document-level features show the same, strong positive sign, suggesting that for robust learning gains and retention of short-term gains, we should optimize strongly towards documents with better coverage of unknown keywords relative to known keywords. We also did find an intuitively strong correlation between total keywords and total words (r=.835, n=283), suggesting that the keyword density of unknown keywords will also likely be a factor positively influencing robust learning outcomes.

Conversely, for Net Retained Knowledge we found a more interesting picture. It was interesting to find that all measures of unknown keyword ratios showed negative but relatively weaker correlations. This makes sense when we consider that Net Retained Knowledge measures not just the retention of previously unknown words but also retention of words that were already known at the time of the pre-test. As such, giving preference to more unknown keywords gave less focus to the participants reinforcing keywords that they may have known only partially at the time, possibly leading to this negative correlation.

We also find that the overall keyword and weighted keyword density measures showed significant and negative correlations at both the document-level and set-level. This suggests that, contrary to what we observed in Section 4, robust retention of knowledge is hurt by providing too many units of knowledge (instances of keyword) in a small amount of text. We also find that Net Retained Knowledge was improved by pages that had more embedded links and those with a lower count of relevant images. This illustrates a tradeoff: whereas these directional features had a negative relationship to short-term learning, they have a positive relationship to long-term retention.

We observe that participants who reported higher ratings for content novelty as an important feature for learning showed significantly better Robust Gains and Retained Gains, suggesting that those who believe more strongly in the importance of content novelty also tend to achieve better Robust Gains and Retained Gains.

## 6 DISCUSSION

We now discuss implications and extensions of our work as it relates to search support of both short- and long-term learning.

**Short-term learning.** We found that short-term measures of vocabulary learning gains are typically improved by: (1) having a lower set-level coverage of unknown keywords versus all keywords; (2) having more contextually relevant images while not having too many total images relative to total word count and (3) having a higher set-level keyword density. We found that we could train robust regression models for predicting learning gains reasonably well with a set of document, user and document-set features.

**Long-term learning.** Due to the nature of the crowd platform used in our experiments, we could not guarantee that we would get

return participants from the original study. However, it turned out that from 600 original responses, we got 81 return participants, as represented by unique (participant ID, topic) tuples that matched against the original dataset. We found evidence that participants who were provided documents chosen exclusively by personalized difficulty-weighted keyword density in the original study showed almost 92% higher Retained Gains of difficult keywords after a nine-month delay compared to those who got documents from a commercial Web search engine.

**IR for Learning.** These results extend the findings from the original study [20] that optimizing purely for difficulty-weighted density improves learning outcomes not only in the short term, but also in the long term. This provides strong support for the utility of efficient, robust document retrieval models to support personalized vocabulary learning at scale.

We also consider some of the limitations of this study. In the study that produced the initial dataset, the authors assumed that a participant's knowledge of a particular term may be modeled as a binary variable (1 if answered correctly and 0 otherwise). More refined and continuous measures of learning would likely give us more accurate knowledge levels that could result in better fitted models. In the robust learning study, we note that the results are based on a relatively small sample size due to the nature of the delayed post-test design. In future work we plan to consider other possible platforms that may be more amenable to more refined longitudinal analysis with a larger study population.

## 7 CONCLUSION

This study analyzed how features of documents and user knowledge related to multiple types of learning outcomes, both short-term and long-term, on a contextual vocabulary learning task. We also presented trained regression models for a variety of learning outcome measures that allowed us to analyze the relative importance of document and user features in predicting learning and retention. We primarily focused on features that could be automatically and quickly computed, to enable these models to be applied at scale in a large variety of possible applications. We also provided a second set of models, specifically trained on non-user features to accommodate realistic scenarios where a user's prior knowledge of an arbitrary topic is not known.

Beyond analyzing short-term learning outcomes, we were able to analyze long-term learning outcomes for a subset of users from the original study who completed a delayed post-test approximately nine months after the initial post-test. Due to the smaller sample size of this subset, we did not provide trained models but we did provide median split analysis of long-term learning outcomes against each feature of the full feature set. Finally, we investigated how different retrieval models were associated with changes to a user's vocabulary knowledge state in the immediate and delayed test stages and found evidence that the personalized retrieval model introduced in [20] provided documents that resulted in almost double the long-term learning gains for higher-difficulty terms compared to corresponding results for a commercial search baseline.

**Acknowledgements.** We thank the anonymous reviewers for their comments. This work was supported in part by the Michigan Institute for Data Science (MIDAS), and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140647 to the University of Michigan. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

## REFERENCES

[1] Mustafa Abualsaud. 2017. *Learning Factors and Determining Document-level Satisfaction In Search-as-Learning.* Master's thesis. University of Waterloo.
[2] Peter Bailey, Liwei Chen, Scott Grosenick, Li Jiang, Yan Li, Paul Reinholdtsen, Charles Salada, Haidong Wang, and Sandy Wong. 2012. User task understanding: a web search engine perspective. In *NII Shonan Meeting on Whole-Session Evaluation of Interactive Information Retrieval Systems, Kanagawa, Japan.*
[3] Elizabeth Ligon Bjork, Jeri L Little, and Benjamin C Storm. 2014. Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition* 3, 3 (2014), 165–170.
[4] Bertram C Brookes. 1980. The foundations of information science Part I. Philosophical aspects. *Journal of Information Science* 2, 3-4 (1980), 125–133.
[5] Kevyn Collins-Thompson and Jamie Callan. 2004. Information Retrieval for Language Tutoring: An Overview of the REAP Project. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04).* ACM, New York, NY, USA, 544–545.
[6] Kevyn Collins-Thompson, Soo Young Rieh, Carl C. Haynes, and Rohail Syed. 2016. Assessing Learning Outcomes in Web Search: A Comparison of Tasks and Query Strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16).* ACM, New York, NY, USA, 163–172.
[7] Cathy De Rosa. 2006. *College students' perceptions of libraries and information resources: A report to the OCLC membership.* OCLC.
[8] Diana DeStefano and Jo-Anne LeFevre. 2007. Cognitive load in hypertext reading: A review. *Computers in Human Behavior* 23, 3 (2007), 1616–1641.
[9] Geoffrey B Duggan and Stephen J Payne. 2008. Knowledge in the head and on the web: Using topic expertise to aid search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 39–48.
[10] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the Journey: A Query Log Analysis of Within-session Learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14).* ACM, New York, NY, USA, 223–232.
[11] Luanne Freund, Rick Kopak, and Heather O'Brien. 2016. The effects of textual environment on reading comprehension: Implications for searching as learning. *Journal of Information Science* 42, 1 (2016), 79–93.
[12] Bernard J Jansen, Danielle Booth, and Brian Smith. 2009. Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management* 45, 6 (2009), 643–663.
[13] Terry Judd and Gregor Kennedy. 2010. A five-year study of on-campus Internet use by undergraduate biomedical students. *Computers & Education* 55, 4 (2010), 1564–1571.
[14] Jin Young Kim, Kevyn Collins-Thompson, Paul N. Bennett, and Susan T. Dumais. 2012. Characterizing Web Content, User Interests, and Search Behavior by Reading Level and Topic. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12).* ACM, New York, NY, USA, 213–222.
[15] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44, 4 (2012), 978–990.
[16] Richard E Mayer. 1997. Multimedia learning: Are we asking the right questions? *Educational Psychologist* 32, 1 (1997), 1–19.
[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
[18] Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. 2016. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science* 42, 1 (2016), 19–34.
[19] Rohail Syed and Kevyn Collins-Thompson. 2017. Optimizing search results for human learning goals. *Information Retrieval Journal* (2017), 1–18.
[20] Rohail Syed and Kevyn Collins-Thompson. 2017. Retrieval algorithms optimized for human learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 555–564.
[21] Manisha Verma, Emine Yilmaz, and Nick Craswell. 2016. On Obtaining Effort Based Judgements for Information Retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining.* ACM, 277–286.
[22] Ryen W White, Susan T Dumais, and Jaime Teevan. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining.* ACM, 132–141.
[23] Barbara M Wildemuth. 2004. The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology* 55, 3 (2004), 246–258.
[24] Wan-Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. 2012. Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proceedings of the 4th Information Interaction in Context Symposium.* ACM, 254–257.
[25] Joerg Zumbach and Maryam Mohraz. 2008. Cognitive load in hypermedia reading comprehension: Influence of text type and linearity. *Computers in Human Behavior* 24, 3 (2008), 875–887.