Modeling the Influence of Format and Depth during Effortful Retrieval Practice

Jaclyn K. Maass
University of Memphis
Institute for Intelligent Systems
Memphis, TN 38152
1-352-584-9103
jkmaass@memphis.edu

Philip I. Pavlik Jr.
University of Memphis
Institute for Intelligent Systems
Memphis, TN 38152
1-901-678-2326
ppavlik@memphis.edu

ABSTRACT

This research combines work in memory, retrieval practice, and depth of processing research. This work aims to identify how the format and depth of a retrieval practice item can be manipulated to increase the effort required to successfully recall or formulate an answer, with the hypothesis that if the effort required to answer an item is increased there will be more benefit to learning. This hypothesis stems from work on desirable difficulties and the effortful retrieval hypothesis. Our data source was an experiment that used a 2 (question depth: factual, applied) x 2 (answer format: multiple choice, short answer) between-subjects design to investigate the effects of these conditions on retrieval practice performance. The experiment was delivered online though Mechanical Turk (n = 178). A logistic regression predicting performance during practice indicates that participants get more (in terms of an increase in future predicted success) from successful retrievals of items that fall within the more difficult level of both the format and depth factors (i.e., short answer and applied). There is also some support that the benefit from multiple choice items may be increased by asking deeper, more applied questions. The application of these results to scheduling effective practice is discussed.

Keywords

Retrieval practice, application, difficulty, multiple choice, short answer, modeling, depth of processing

1. INTRODUCTION

The testing effect is the well-replicated benefit of retrieval practice (i.e., “testing yourself”), typically over the course of several repetitions [e.g., 1; 7; 16; 30; 33]. Experiments often compare the benefit of active retrieval against re-reading or re-studying written material and much of the early work in this field utilized a more traditional cognitive psychology experimental setup (e.g., using word lists/pairs or isolated facts, controlling for prior knowledge, and post testing with verbatim items repeated from practice). This design, however, does not well represent how retrieval practice would be implemented in authentic educational settings. For implementation in classrooms, issues that have real-world importance to educators, such as the format of the questions and the ease of administration, should be considered.

The effect of answer format has long been of interest not only to educational researchers (e.g., comparing multiple choice, fil-in-the-blank, essays, etc.), but also to cognitive psychologists (e.g., comparing recognition, cued or free recall, etc.). Research has shown a continuum in terms of performance/difficulty ranging from recognition, to cued recall, to free recall which translates roughly in educational terms to multiple choice, short answer, and essay questions. This ordering is found consistently in research and is summed up nicely by Glover’s [13] work which reported the effectiveness of three formats used during retrieval practice (referred to as intervening tests): free recall, cued recall, and recognition (see Experiment 4). After reading a passage and having intervening tests in one of the three formats, participants took a retention test after four days. The free recall intervening test was an open-ended format, with participants writing what they remembered from the passage. The cued recall intervening test was a fill-in-the-blank format, using sentences paraphrased from the text. The recognition intervening tests required the participants to identify which of several sentences they had read previously in the text. The final retention test included items in each of the three formats (across the posttests in Experiments 4a, 4b, and 4c). A very clear pattern emerged: the fewer cues there were available during practice (e.g. free recall provided the fewest cues), the better participants performed on the final retention test. Those who had intervening tests in a free recall format out-performed participants in the cued recall condition on the final retention test (statistically significant difference), who in turn outperformed those who practiced with a recognition task (not statistically significant). Perhaps most importantly, this advantage held regardless of the format of the retention test, which included all three formats [13].

There are several other studies which show us the benefit of using fewer cues (e.g., short answer format) during retrieval practice. Kang, McDermott, and Roediger III [18] had participants read several journal articles. After reading each article, participants completed one of four possible tasks- a multiple choice test, a short answer test, reading relevant facts from the text, or a questionnaire (i.e., filler task). When feedback was provided during the practice tests, those items that had been practiced in short answer format had significantly higher scores on the final test. Results also indicated that practice with multiple choice testing was no better than re-reading relevant facts. The researchers concluded with a recommendation for practice testing with short answer items. Similar results were found in work by McDaniel, Anderson, Derbish, and Morrisette [22], which indicated that weekly practice tests were more effective in increasing final test performance when the weekly practice was in the form of short answer questions compared to multiple choice items. Since the final test was only in multiple choice format, it suggests another benefit of short answer practice is the ability to overcome transfer-appropriate-processing effects, which would predict that the final test performance would be highest when it matched the conditions of earlier practice [24]. In other words, short answer may be a better alternative to multiple choice regardless of how you assess it.

One possible reason for why practice with short answer often outperforms multiple choice on final outcome measures is the amount of effort required for retrieval [18]. This general benefit of effortful retrieval has been referred to as the retrieval effort hypothesis, which was motivated by Bjork's [4; 5] desirable difficulty framework and Craik and Lockhart's [11] depth of processing research. The retrieval effort hypothesis, as defined by Pyc and Rawson [29], claims that there is more memorial benefit from successful retrieval practice when it is difficult than when it is less difficult. This follows from the desirable difficulty framework, which suggests that practice which is made more difficult (up to a certain point) will lead to more durable and generalizable learning [4]. The desirable difficulty framework sets a theoretical upper bound on the level of difficulty appropriate for effective learning, which can depend on several individual differences including prior knowledge and working memory capacity. This is similar to the assistance dilemma [19], which suggests there is an optimal middle-ground in terms of how difficult a task should be, and/or how much assistance should be offered to a student during a learning task.

The goal of the current work was to generate data to further investigate the effect of effortful retrieval practice, and specifically, how we can equate the effort required to successfully answer multiple choice items with the effort required for short answer items. One way to address this is to increase the effort required to correctly answer a multiple choice question, and the way to do so may lie within the depth of processing required to respond to the question. By asking a deeper, more applied question, rather than the more common text-based factual question, perhaps we can encourage deeper processing so as to increase the effort required for multiple choice questions.

The depth of processing framework suggests that information which is processed on a deeper level will be encoded in a more elaborate and durable manner, with depth referring to greater semantic or cognitive processing [11]. Craik [10] further defines depth as "the qualitative type of processing carried out on the stimulus…" (p. 307). Questions that require more cognitive processing to successfully answer have also been referred to as deep-reasoning questions. Deep-reasoning questions rely on a student's logic and reasoning abilities and are thought to tap into more complete and coherent understanding [14]. Deep-reasoning questions are embedded in the deeper levels of cognition in Bloom's [6] taxonomy, and both have been shown to be positively correlated with final examination scores [14]. In the current work we attempt to increase the difficulty of multiple choice items by asking deeper, more applied questions, and mine our data to compare the benefit that we see from these more difficult multiple choice items with typical benefit from asking factual short answer items.

The interaction of answer format and depth of processing has been investigated to some degree in work by Smith and Karpicke [31], which compared three answer format conditions :multiple choice, short answer, and hybrid conditions which consisted of short answer-multiple choice pairings. Question type during retrieval practice (i.e., factual and inference questions) was a within-subjects factor (Experiments 1, 2, and 3), but this factor was collapsed in the analyses of final assessment performance. They concluded that practice with short answer could lead to higher performance on the final assessment (compared to practice with multiple choice questions), if students achieve a high proportion of correct short answer responses during practice. Smith and Karpicke therefore attempted to equate the initial practice performance between the short answer and multiple choice conditions. Those results are discussed in more detail in their paper [31], but of importance to the current work is that they attempted to raise performance on short answer questions up to the performance on multiple choice items. The current work will essentially attempt the opposite- increasing the difficulty (or lowering the performance) of multiple choice in an attempt to "equate" it to short answer. Therefore, the design of the current data collection was partially inspired by that of Smith and Karpicke, in an attempt to get more fine-grained information about the interaction between format and depth during practice, and their effect on different format and depths at posttest.

In theory, the multiple choice questions in Smith and Karpicke's work were more difficult when the multiple choice was an inference item, rather than factual, but the nature of their inference questions appears to be fairly straightforward, without requiring much more effort than the factual questions. Specifically, the inference items required participants to combine different facts they had previously read in order to draw a conclusion/answer that had not been explicit in the text. However, for most (if not all) of the inference items, the facts required to answer them were presented within a single paragraph. This is not inherently problematic, but it is important to take note of if your objective is to increase the effort required to answer a multiple choice item, since it brings into question the level of difficulty of the inference questions. For example, an inference would be more difficult to make if it required retrieving and combining more than two facts, or if those facts were presented further apart from each other in the text. Further, the answer options in Smith and Karpicke's multiple choice items only included a single option that appeared in the text- the correct answer option. Thus, these questions become purely a measure of memory (of a previously read text), rather than understanding or learning. In other words, the students wind up asking themselves, "Which of these answer options did I see in/ matches with the text I read earlier" rather than, "Which of these options make sense and accurately reflects what I read?" This only serves to further reduce the difficulty of multiple choice practice. To alleviate this, the multiple choice answer options for the current work were all feasible, text related answers that underwent several iterations, described in detail in the materials section.

1.1 The Current Study

The current study focuses on two ways to increase the difficulty of retrieval: through the amount of retrieval cues available (i.e., the answer format: multiple choice or short answer) and through the depth of processing required to successfully answer the question itself (i.e., the question depth: factual or applied). We attempt to mine our data to determine whether or not the difficulty of multiple choice be increased by asking a deeper question, and whether difficulty created through varying amounts of retrieval cues (i.e., the answer format) is similar to the difficulty created through the depth of the question.

The purpose of this paper is to investigate the effect of question format, depth, and individual differences during retrieval practice. Although the experiment tested several types of transfer at the posttest (e.g., format, depth, and unpracticed information), this paper is predominantly focused on dissecting the mechanisms at play during practice. In order to do this, we employed a method of model-based discovery [3] in which previously developed models are adapted to fit the particular research questions and data being mined. In order to create a more complete picture, however, some descriptive information regarding posttest performance is provided, although it is not the main focus of this paper.

2. METHODS

2.1 Design

The experiment manipulated difficulty of retrieval practice through a 2 (question depth: factual, applied) x 2 (answer format: multiple choice, short answer) between-subjects design. The difficulty of the posttest was also manipulated with a 2 (posttest question depth: factual, applied) x 2 (posttest answer format: multiple choice, short answer) x 2 (concepts: practiced, unpracticed) fully factorial within-subjects design. This resulted in four between-subjects retrieval practice conditions (Factual MC, Applied MC, Factual SA, or Applied SA), and posttest questions in each of those four conditions, allowing for measures of transfer to a different depth and format, as well as transfer to previously unpracticed concepts. Prior knowledge was assessed by a 6-item pretest on factual questions, half randomly assigned per participant to multiple choice and half to short answer format. This experiment did not include a control condition with no retrieval practice. This was a conscious decision since the testing effect is widely accepted as a reliable phenomenon, and the current design allows for a more tractable, and fine-grained investigation of specific components of retrieval practice.

2.2 Participants

One hundred ninety-three participants were recruited through the Mechanical Turk (MTurk) online data collection platform. The only requirements were for the participants to be at least 18 years of age, a native English speaker, from the United States or Canada, and be a reliable MTurk worker. The last requirement was defined as a worker who had completed at least 50 MTurk tasks with at least a 95% approval rate. Data for 10 participants were removed due to the participants having ten or more time-outs during the experiment and five participants' data were removed due to glitches in the system (*n*=178, 58% male). Within this sample, 45% were in the age range of 26-34 years, 31% were in the age range of 35-54 years, 30% were between 18-25 years, and 4% were between 55-64 years. Most participants reported that their highest level of completed education was "Some college" (37.2%), followed by "High school/ GED" (17.7%), "Graduate degree" (6.6%), and "Less than high school" (<1%). Each MTurk worker was paid $5.00 for participation

2.3 Materials[1]

2.3.1 Text

The experimental text was 995 words in length and pertained to the circulatory system. It was compiled from texts used in previous research [15; 35], and is estimated to be at a Flesch-Kincaid 6th grade reading level (https://readability-score.com).

2.3.2 Factual and Applied Items

Sixteen concepts were extracted from the text to be used for the creation of factual and applied questions. These concepts represent what we believe to be the crucial components in the text, and are aligned with, and expanded from, the factual questions previously used with these materials [23; 35]. The first author, along with another graduate student familiar with this line of research, created a factual and an applied question based on each of the 16 key concepts. The factual versions for the 16 concepts are taken directly from the text. For example, the text states, "The heart is a pump. Its walls are made of thick muscle. They can squeeze (contract) to send blood rushing out." The factual question for this concept asks, "Which component of the circulatory system acts as a pump?" Answer: the heart.

For each of the 16 concepts, we also created an applied question through brainstorming sessions by asking ourselves the questions, "Why is this fact or component important to the circulatory system?" or "What would happen if this component was not functioning properly?" In most of these cases, the 16 applied questions reference the consequence of the factual relationship (described in the text) not holding true. For example, many applied questions require participants to predict outcomes given a certain component not functioning normally. The key principle for the applied questions is that participants must retrieve the necessary fact or facts from memory (presented previously in the text) and apply them in a new way. Importantly, the text only discusses the normally functioning circulatory system, and presents the material at the factual level, without much elaboration. Therefore, the applied questions are not presented explicitly in the text, but can be answered by processing and recombining the facts contained within the text. For the previous example, the concept of the heart acting as a pump, the applied question is, "Why doesn't oxygen rich blood flow directly from the lungs to the rest of the body?" Answer: Because blood requires a pump, the heart, to push it through the body.

2.3.3 Multiple Choice Answer Options

Each question, both factual and applied, required three (incorrect) answer options for the multiple choice format. The incorrect answer options were created based on common misconceptions about the circulatory system. Information on misconceptions was gathered through past research [e.g., 32] and pilot testing (common incorrect responses to the questions in short answer format). Once three answer options (in addition to the correct answer) were created for each of the factual and applied questions, additional pilot testing confirmed that the frequencies of responses for each of the three incorrect answer choices were not substantially different from each other. This method for creating the answer options was specifically done in an attempt to not lessen the effort required to answer a multiple choice item by using answer options that were unrelated or too easy for a participant to exclude as a possible answer.

2.4 Procedure

The experiment consisted of four portions (pretest, reading, retrieval practice, and posttest) within a single session delivered online through Amazon's Mechanical Turk web service using the MoFaCTS online tutoring system (http://mofacts.optimallearning.org/) [27]. The entire experiment took an average of approximately 60 minutes for participants to complete. After obtaining informed consent, participants completed a pretest consisting of six factual questions. For each participant, half of the questions were randomly assigned to short answer format and the other half to multiple choice. These six questions were created from the text in the same way as those for retrieval practice, but did not overlap with the 16 concepts covered in retrieval practice to reduce the possibility of priming. No corrective feedback was given during the pretest.

[1] Experimental materials are available upon request; please contact the first author.

Next, the participants were asked to read the Circulatory System text which was displayed on a single screen (with a scroll bar). For this portion, participants were instructed to not take notes while they read the text. Participants read at their own pace without a time limit. The average time spent reading was approximately seven minutes.

Following the reading portion, participants began retrieval practice. Each participant was randomly assigned to practice with either factual MC, applied MC, factual SA, or applied SA questions. Retrieval practice consisted of eight questions (each representing a different concept covered in the text), repeated four times each. These eight items were randomly selected for each participant from the list of 16 concepts. The order of the eight questions was randomized for each of the four "blocks" of repetition. Corrective feedback was given immediately after participants entered their responses. Correct responses allowed the participant to immediately move on to the next item; incorrect responses were followed by a review period of 10 seconds, during which the correct response was shown on the screen. This feedback procedure not only provided the correct answer for the participant to review, but also provided an incentive for participants to try their best, since correct answers allowed the participant to "skip" the mandatory 10-second review period. In other words, participants would quickly realize that random guessing or poor effort would only increase the length of the experiment.

The final portion of the experiment was the posttest, which was given after a delay of approximately one minute. During this delay the participants were instructed to complete a "current emotion" survey discussed below. The eight concepts studied during retrieval practice were included in the posttest, but each was randomly assigned to be tested in one of the four format/depth conditions. Each participant also answered an additional eight posttest items (two in each of the format/depth conditions) which reference the eight remaining concepts that were not randomly selected for retrieval practice. This allowed us to see how well each practice condition transferred to similar but previously untested material. Each of the 16 posttest questions were presented once, in random order, without corrective feedback.

At three different points in the experiment, participants responded to a set of six "current emotion" questions. The three time-points were: before the retrieval practice to obtain a baseline, immediately after retrieval practice to look for an effect of practice condition on affect, and immediately after posttest to determine if the change in format and depth at posttest had an adverse effect on affect. Specifically, participants were asked to rate on a scale of 1 (Strongly Disagree) to 5 (Strongly Agree) how much they agree with the statement, "Currently, I am feeling _____." This question was asked six times, with a different affect provided in the blank. The six affects were: anxious, bored, confused, discouraged, frustrated, and unfocused/distracted. Demographic information was also collected at the conclusion of the experiment.

2.5 Scoring

All questions were scored immediately by the system and received a score of 1 or 0 (although this value was not explicitly displayed to the participant). MoFaCTS (the online drill-trial problem authoring and deployment platform we used) scored short answer items by matching words in the participants' responses to key terms necessary to answer the question correctly. Pilot testing revealed common (acceptable) synonyms and alternative words that we incorporated into the system to allow for slight variation in what was considered a correct response. For example, the (complete) correct answer for the (factual) question, "Where is the heart located in relation to the lungs?" is "The heart is located between the lungs." The system scored the responses to this item based on whether or not it contained the word "between" or "middle." The use of regular expressions embedded in the MoFaCTS programming allowed for any of the following responses to be counted as correct: "between the lungs", "the heart is between the lungs", or "the heart's in the middle of the lungs." The regular expressions in the system also accounted for ordering when applicable; for example, ordering is essential for the (factual) question, "Which gas do the cells of the body require to function and which gas do they expel as waste?" Participants received corrective feedback (either "Correct" or "Incorrect. The correct response is _______") after each item in the retrieval practice portion, but not during the pretest or posttest.

3. RESULTS AND DISCUSSION

3.1 Overall Performance

Before we discuss the results of mining our retrieval practice data, it may be helpful to review the broader results of the experiment. Table 1 provides an overview of the average scores for the practice (8 items with 4 trials each), the portion of the posttest containing the eight concepts previously practiced, each randomly assigned to one of the four format/depth conditions, (total of 8 trials), and the portion of the posttest which consisted of eight previously unpracticed concepts, each randomly assigned to one of the four format/depth conditions (total of 8 trials).

The average performances during retrieval practice, provided in Table 1, support the general ordering of performance we expected for each condition. Namely, the Factual MC condition was the least difficult, with the highest performance during practice, the Applied SA was the most difficult condition as indicated by the lowest performance during practice, and the Applied MC and Factual SA fall in between in terms of performance during practice. A between-subjects Analysis of Variance (ANOVA) indicated significant differences between the four conditions, F (3,174) = 28.49, p<.001. Post hoc pairwise comparisons indicate that the only two conditions that are *not* significantly different from each other are the Applied MC and Factual SA conditions (p = .19). All other conditions are significantly different from each other (all p's < .05).

Table 1. Means and Standard Deviations for Practice and Posttest Performance by Condition

Retrieval Practice Conditions	Average Practice Performance	Average Posttest Scores[†]	
		Practiced Concepts	Unpracticed Concepts
Factual MC (n = 46)	.85 (.12)	.65 (.17)	.45 (.23)
Applied MC (n = 42)	.77 (.17)	.70 (.21)	.45 (.24)
Factual SA (n = 47)	.73 (.15)	.69 (.14)	.50 (.19)
Applied SA (n = 43)	.55 (.18)	.68 (.20)	.47 (.21)

Note: [†] collapsed across all format/depth posttest conditions. Standard deviations in parentheses.

Table 1 also displays posttest performance for each condition on the eight concepts they had been tested on during practice, as well as on eight concepts they had read about in the text, but had not actively practiced. Between-subjects ANOVA's showed no significant differences between conditions for performance on either posttest. Note that the drop in performance from practice to the practiced concepts posttest is due to the within-subjects nature of the posttest conditions. In other words, the eight concepts were only practiced in one condition, but were then randomly assigned to be tested in one of the four depth/format conditions in the posttest, meaning that participants had two items in the posttest of practiced concepts that were in a different format, two that were in a different depth, and two that were in a different format and depth. These different types of transfer in the posttest for the practiced and unpracticed concepts therefore resulted in lowered overall performance. Although not significant, we do see that the Factual MC condition was most affected by these transfer items for the posttest on practiced concepts.

While the ANOVA's offer us a broad view of overall performance, in order to truly answer our research questions we will need a finer-grained analysis. Mining our data and creating a model of learning will give us a more in depth look at what is taking place during retrieval practice.

3.2 Modeling Retrieval Practice

A logistic mixed-effects regression was created to model learning during retrieval practice. Since retrieval practice conditions differed in the question depth and answer format factors according to the result above, this model is meant to dissect the differential learning caused by each type of question. The model is based on a Performance Factors Analysis (PFA) where performance is predicted on subsequent trials as a function of the performance on prior trials [26]. Unlike Additive Factors Modeling (AFM) [8], PFA captures prior performance by two parameters, differentiating the effect of prior incorrect (unsuccessful) and correct (successful) trials. We chose to use PFA to separate these components because it would allow us to look into the difference in predictive ability between successful versus unsuccessful prior retrievals. This comparison would indicate if the benefit of retrieval practice is dependent on successful retrieval, or if the mere attempt at retrieval (i.e., incorrect trials) also results in better performance.

Modeling the data included several iterations guided by our hypotheses concerning the effects of format, depth, and prior knowledge. We began with the basic components of a PFA model: two parameters to capture the count of prior correct and incorrect trials. We also included pretest score and a random effect of subject, all of which were significant.

We then added in features we suspected would affect performance based on the cognitive and educational research discussed above, namely, the format and depth of the practiced items. We used one parameter to capture the format of the current item and one to capture the depth of the current item. We also tried adding measures of response time (e.g., time spent reading the text prior to practice, average time spent on all previous trials, and average time spent on previous trials with the specific item, etc.) but none were significant in the model. Next, we added interactions between all factors that had proven significant at that point (e.g., count of prior correct by depth, count of prior incorrect by pretest, depth by format, etc.) Only two of these interactions were significant: count of prior correct by format and count of prior correct by depth, which were retained in the final model. Finally, several measures of affect were added to the model (i.e., the affective score).

The final additions to the model included measures of affect. Remember that our measure of affect consisted of six questions which each used a 5-point Likert-item (1- Strongly Disagree to 5-Strongly Agree) for participants to rate how much they agreed with the statement: "Currently I am feeling _____" for each of the six different affects (anxious, bored, confused, discouraged, frustrated, and unfocused/distracted). Ratings for each of these six affects were collected before and after retrieval practice (and after posttest, but that was not relevant to modeling the learning during practice). We tested the model using six parameters of the affects before practices, and then six parameters to capture the affect after practice. We decided to try to approximate participants affective states during practice by averaging the self-reported levels of affect reported before and after practice. It should be noted that there was not much change in affect from before to after retrieval practice, and each of the three measures (the "before" ratings, the "after" ratings, and the average of the two) performed similarly in the model. Confusion (averaged to capture affect during practice) was the only affect factor that improved the fit of the model. The last step was adding in interactions between this confusion measure and the count or prior correct and incorrect trials, of which only the latter was significant. The final model, summarized in Table 2, retained each of the parameters that achieved significance throughout our modeling process.

The final model had an R^2 of .359, with 5,696 total observations from 178 participants. The *AIC* was 4838.2, the *BIC* was 4904.6, and the Log Likelihood was 4818.2. Table 2 summarizes the fixed effects parameter values of the final model. Not included in Table 2 is the random effect of Participant (*SD* = 0.669). For the format and depth parameters, a value of 0 was assigned to the less difficult level (i.e., MC and Factual) and a value of 1 was assigned to the more difficult level of each factor (i.e., SA and Applied). For each of the parameters involving the count of prior correct or incorrect trials, the log of (1 + the prior count) was taken to account for diminishing marginal returns expected from the power law of practice [25]. Figure 1 also illustrates the fit of the model (left) to the participants' data (right).

Ten runs of a 10-fold cross-validation revealed that the model retained validity when comparing the training folds (R^2 = .336) to the testing folds (R^2 = .329). The CV proportion (training folds R^2 divided by testing folds R^2) for the model indicated that 97.9% of the validity of the model was retained in the held out data.

Table 2. Summary of Fixed Effects for Logistic Regression Model Predicting Future Success

Parameter	Parameter Estimate	*SE*	*Z*-value
Intercept	-0.11	.19	-.56
Pretest	1.95	.30	6.50
Count of Prior Correct	1.82	.16	11.72
Count of Prior Incorrect	1.47	.15	9.88
Format	-1.22	.14	-8.93
Depth	-0.82	.13	-6.06
Prior Correct x Format	1.13	.19	5.93
Prior Correct x Depth	0.36†	.19	1.93
Prior Incorrect x Confusion	-0.18	.05	-3.78

Note: † p < .05; all other parameters are significant at the p < .001 level. For the Format and Depth parameters, MC and factual are coded as 0, and SA and applied are coded as 1, respectively.

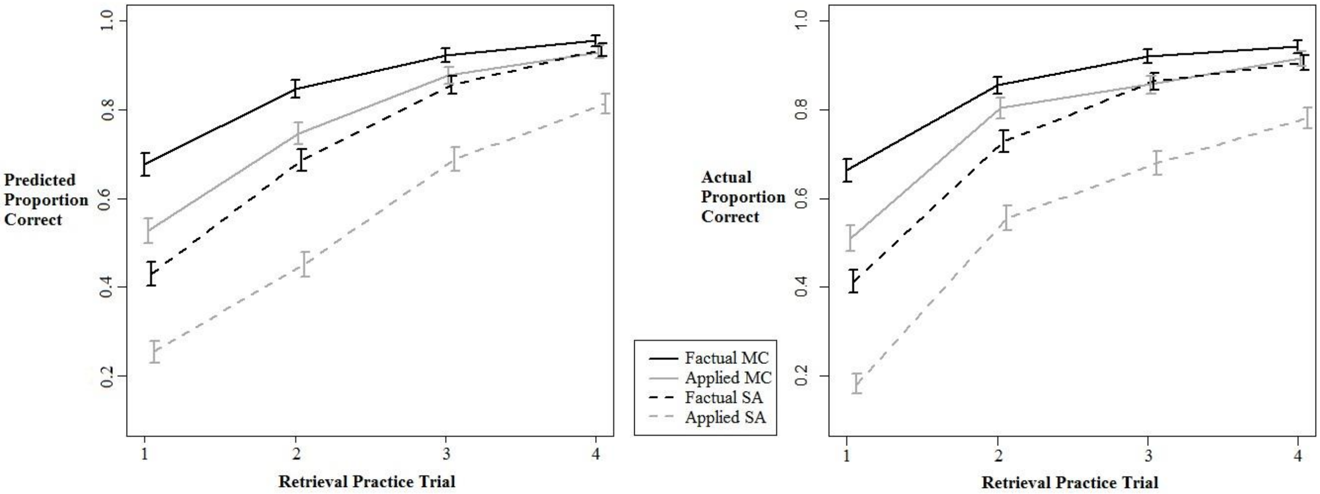

Figure 1. Side by side comparison of the model's predicted performance (left) and the participants' actual performance (right) during the four trials of retrieval practice.

3.3 Model Interpretation

One of the first things the data mining reveals is that correct retrieval (specifically recall) is important for learning. However, the current model also indicates a benefit from unsuccessful retrieval, although to a smaller degree. It is worth noting the model also shows a (lesser) benefit from unsuccessful trials. When comparing just the effect of prior correct and incorrect practice trials, it appears that they offer almost equivalent additions to the prediction/model (1.47 vs 1.82). However, the count of prior correct also interacts with the depth and with the format. For three out of the four practice conditions, these increase the predictive ability of previous successful practices. Therefore, taken all-together, there is much more of a positive effect of previous correct trials than incorrect trials. For example, in the Applied SA condition with one previous correct trial and one previous incorrect trial, successful practices is more than twice as impactful on future performance as previous unsuccessful practices when taking the interactions into account. This difference between the influence from previous correct versus incorrect trials is made even greater if the student has a higher level of confusion (as indicated by the negative estimate for the confusion*incorrect count parameter). This result adds to the building body of research that suggests it is successful retrieval, and not just the attempt to retrieve, that is beneficial to learning [20; 21; 29]. Thus, when it comes to supplying challenging questions for retrieval practice, we must be sure that the questions are at an appropriate difficulty-level for the student, so the student can be successful enough to gain from such practice.

Our model also shows how the format and depth of a practice item influence performance. First we see that the average performance for multiple choice practice is significantly higher than practice with short answer (as indicated by the overall performance of the multiple choice conditions during practice in Table 1 and the -1.22 estimate for short answer practice in Table 2 and), which indicates that multiple choice is the better option in terms of allowing for a higher percentage of successful practice. However, we also saw in the model above that there is more gained from successful short answer practice than is gained from successful multiple choice practice (the Prior Correct x Format parameter). This result are aligned with prior work which suggests that the short answer format may not be universally "better," especially if students are not getting a sufficient amount of those questions correct [31]. Based on these results, it is reasonable to suggest that in order to schedule effective practice, students should be given questions that have a higher likelihood of being answering correctly. If we assume that for the most part, students have a lower level of prior knowledge at the beginning of practice/learning a topic, multiple choice item may permit learning by boosting success. However, since successful short answer practice offers more of a benefit (than multiple choice), it seems that students should eventually transition into short answer practice as they become more proficient. In other words, practice should begin with the less effortful item-type and transition to the more effortful (and more beneficial) item once students reach some level of mastery.

The same may be said for practice with the deeper applied items, over the more text-based factual questions, in that students will get the factual items correct more often, but there is more gained from successful applied practice than from successful factual practice. Again, students might benefit most from beginning with the easier depth (factual/ text-based) and finishing retrieval practice with more difficult, applied questions. The goal it seems, should be to get students to a point where they can get many successful retrieval attempts with SA and/or applied items. This suggestion aligns with ideas in several areas of education research including scaffolding [17], zone of proximal development, and concreteness fading [34]. Determining the optimal level of mastery is an important component though, since increased redundancy during learning (repeated practice of known information) has been shown to offer decreasing marginal returns [9; 28]. Our model also illustrates the importance of taking prior knowledge into account when designing tutoring systems and practice schedules. Some students might be able to begin right away with more difficult items (e.g., applied short answer) and others would benefit from beginning practice with factual multiple choice questions and progress from there.

3.3.1 Affect in the Model

The work concerning affect in the current study is exploratory in nature and was meant to give us an indication of which affective states might be the most important to investigate further in future experiments. Our measure of affective states indicated that the most influential affect was confusion. The interaction between the count of prior unsuccessful trials and self-reported confusion level in our model shows that when a learner answers more questions incorrectly, higher confusion predicts a much larger negative effect than if a learner has higher confusion but is still having mostly successful practice. This preliminary result appears to align with previous findings which suggest that confusion can be an important component during learning, and is beneficial when students identify that confusion and work to clarify it (i.e. start to produce correct responses), but detrimental when the confusion is overwhelming or the student fails to remedy it [12].

Unlike previous work by Baker, et. al., [2] we did not find any significant impact of frustration or boredom (nor for the other affective states we asked participants about: anxiousness, discouragement, and distractedness). As the current work was meant to serve only as an exploration of affect during retrieval practice, this is an area that we may investigate further in the future. In future work we may implement pop-up/immediate questions concerning the participant's current affective, or specifically their level of confusion, after more than one incorrect response to measure affect/ changes in confusion during bouts of unsuccessful practice.

3.4 General Conclusions

Our model of performance during retrieval practice indicates a benefit for successful retrieval of short answer over multiple choice items. Likewise, there is a benefit from successful retrieval of applied items over factual items which supports the effortful retrieval hypothesis, that successful trials with more difficult items are better than success on less difficult items. Our hypothesis that the difficulty of multiple choice items could be increased (and equated with difficulty of factual short answer items) by asking applied questions, could potentially be supported by the non-significant difference in practice performances, although more analyses will be necessary before making this conclusion. However, format appears to be a more powerful predictor of future success than depth. This may suggest that the difficulty of retrieving information from memory created from less cues (short answer items), is more beneficial than difficulty created through the effortful processing and reasoning with retrieved information (applied items). We recognize that the construct of retrieval effort could be considered too broad of an explanation for our results. While retrieval effort may not capture all the nuances involved in understanding retrieval, we believe it offers a parsimonious general framework under which several mechanisms are captured. Understanding the role that effort plays in retrieval practice will benefit from future work that investigates the differences in more fine-grained mechanisms such as individual difference in strategy use and/or cognitive processes involved in practice with each question type.

4. ACKNOWLEDGMENTS

Our thanks to the University of Memphis's Institute for Intelligent Systems for providing the resources and support necessary to conduct this research.

5. REFERENCES

[1] Agarwal, P.K., Karpicke, J.D., Kang, S.H., Roediger III, H.L., and McDermott, K.B., 2008. Examining the testing effect with open-and closed-book tests. *Applied Cognitive Psychology 22*, 7, 861-876.

[2] Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., and Graesser, A.C., 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies 68*, 4, 223-241. DOI= http://dx.doi.org/10.1016/j.ijhcs.2009.12.003.

[3] Baker, R.S.J.d. and Yacef, K., 2009. The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining 1*, 1, 3-17.

[4] Bjork, R.A., 1994. Memory and metamemory considerations in the training of human beings. In *Metacognition: Knowing about knowing*, J.M.A.P. Shimamura Ed. The MIT Press, Cambridge, MA, US, 185-205.

[5] Bjork, R.A., 1999. Assessing our own competence: Heuristics and illusions. In *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application*, D.G.A. Koriat Ed. The MIT Press, Cambridge, MA, US, 435-459.

[6] Bloom, B.S., 1956. *Taxonomy of Educational Objectives: The Classification of Education Goals. Cognitive Domain. Handbook 1*. Longman.

[7] Carrier, M. and Pashler, H., 1992. The influence of retrieval on retention. *Memory & Cognition 20*, 6 (Nov), 633-642.

[8] Cen, H., Koedinger, K., and Junker, B., 2008. Comparing Two IRT Models for Conjunctive Skills. In *Intelligent Tutoring Systems: 9th International Conference, ITS 2008, Montreal, Canada, June 23-27, 2008 Proceedings*, B.P. Woolf, E. Aïmeur, R. Nkambou and S. Lajoie Eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 796-798. DOI= http://dx.doi.org/10.1007/978-3-540-69132-7_111.

[9] Cen, H., Koedinger, K.R., and Junker, B., 2007. Is Over Practice Necessary? – Improving Learning Efficiency with the Cognitive Tutor through Educational Data Mining. In *Proceedings of the 13th International Conference on Artificial Intelligence in Education* (Los Angeles, CA2007).

[10] Craik, F.I., 2002. Levels of processing: Past, present . . . and future? *Memory 10*, 5-6 (Sep), 305-318.

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

[12] D'Mello, S., Lehman, B., Pekrun, R., and Graesser, A., 2014. Confusion can be beneficial for learning. *Learning and Instruction 29*, 153-170.

[13] Glover, J.A., 1989. The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology 81*, 3 (Sep), 392-399.

[14] Graesser, A.C. and Person, N.K., 1994. Question asking during tutoring. *American Educational Research Journal 31*, 1, 104-137.

[15] Hathorn, L.G. and Rawson, K.A., 2012. The roles of embedded monitoring requests and questions in improving mental models of computer-based scientific text. *Computers & Education 59*, 3, 1021-1031. DOI= http://dx.doi.org/10.1016/j.compedu.2012.04.014.

[16] Johnson, C.I. and Mayer, R.E., 2009. A testing effect with multimedia learning. *Journal of Educational Psychology 101*, 3, 621.

[17] Kang, H., Thompson, J., and Windschitl, M., 2014. Creating Opportunities for Students to Show What They Know: The Role of Scaffolding in Assessment Tasks. *Science Education 98*, 4, 674-704. DOI= http://dx.doi.org/10.1002/sce.21123.

[18] Kang, S.H., McDermott, K.B., and Roediger III, H.L., 2007. Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology 19*, 4-5, 528-558.

[19] Koedinger, K.R., Pavlik Jr., P.I., McLaren, B.M., and Aleven, V., 2008. Is it Better to Give than to Receive? The

Assistance Dilemma as a Fundamental Unsolved Problem in the Cognitive Science of Learning and Instruction. In *Proceedings of the 30th Conference of the Cognitive Science Society*, V. Sloutsky, B. Love and K. McRae Eds., Washington, D.C., 2155-2160.

[20] Maass, J.K. and Pavlik Jr, P.I., 2013. Using learner modeling to determine effective conditions of learning for optimal transfer. In *Artificial Intelligence in Education* Springer, 189-198.

[21] Maass, J.K., Pavlik Jr, P.I., and Hua, H., 2015. How Spacing and Variable Retrieval Practice Affect the Learning of Statistics Concepts. In *Proceedings of the 17th International Artificial Intelligence in Education Conference* Springer-Verlag, Berlin, Heidelberg.

[22] McDaniel, M.A., Anderson, J.L., Derbish, M.H., and Morrisette, N., 2007. Testing the testing effect in the classroom. *European Journal of Cognitive Psychology 19*, 4-5, 494-513.

[23] Michael, A.L., Klee, T., Bransford, J.D., and Warren, S.F., 1993. The transition form theory to therapy: Test of two instructional methods. *Applied Cognitive Psychology 7*, 2, 139-153. DOI= http://dx.doi.org/10.1002/acp.2350070206.

[24] Morris, C.D., Bransford, J.D., and Franks, J.J., 1977. Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior 16*, 5, 519-533.

[25] Newell, A. and Rosenbloom, P.S., 1981. Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition 1*, 1-55.

[26] Pavlik Jr., P.I., Cen, H., and Koedinger, K.R., 2009. Performance Factors Analysis -- A New Alternative to Knowledge Tracing. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, V. Dimitrova, R. Mizoguchi, B.d. Boulay and A. Graesser Eds., Brighton, England, 531-538.

[27] Pavlik Jr., P.I., Kelly, C., and Maass, J.K., 2016 submitted. Using the Mobile Fact and Concept Training System (MoFaCTS).

[28] Pavlik Jr., P.I., Maass, J.K., and Hua, H., 2015, November. Redundancy causes spacing effects. In *56th Annual Meeting of the Psychonomic Society*, Chicago.

[29] Pyc, M.A. and Rawson, K.A., 2009. Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language 60*, 4, 437-447.

[30] Roediger III, H.L. and Karpicke, J.D., 2006. The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science 1*, 3, 181-210. DOI= http://dx.doi.org/10.2307/40212166.

[31] Smith, M.A. and Karpicke, J.D., 2014. Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory 22*, 7, 784-802.

[32] Sungur, S., Tekkaya, C., and Geban, Ö., 2001. The contribution of conceptual change texts accompanied by concept mapping to students' understanding of the human circulatory system. *School Science and Mathematics 101*, 2, 91-101.

[33] Thompson, C.P., Wenger, S.K., and Bartling, C.A., 1978. How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning & Memory 4*, 3 (May), 210-221.

[34] Tullis, J.G., Goldstone, R.L., and Hanson, A.J., 2015. Scheduling Scaffolding: The Extent and Arrangement of Assistance During Training Impacts Test Performance. *Journal of Motor Behavior 47*, 5 (2015/09/03), 442-452. DOI= http://dx.doi.org/10.1080/00222895.2015.1008686.

[35] Wolfe, M.B., Schreiner, M., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W., and Landauer, T.K., 1998. Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes 25*, 2-3, 309-336.