Short-Answer Responses to STEM Exercises: Measuring Response Validity and Its Impact on Learning

Andrew Waters
Rice University / OpenStax
Houston, TX
aew2@rice.edu

Phillip Grimaldi
Rice University / OpenStax
Houston, TX
pjg3@rice.edu

Andrew Lan
Princeton University
Princeton, NJ
andrew.lan@princeton.edu

Richard Baraniuk
Rice University
Houston, TX
richb@rice.edu

ABSTRACT

Retrieval practice is a study technique in which students answer questions related to target material, and has been demonstrated to be an effective way to promote learning and retention. Educational technology commonly leverages multiple-choice questions for retrieval practice, but short-answer questions hold the potential to provide better learning outcomes. Unfortunately, students in online settings often exhibit little effort when crafting short-answer responses. Instead, students often produce invalid (or garbage) responses that are off-topic and do not relate to the question being answered. In this study, we consider the effect of response validity on retrieval practice. To do this, we first develop GarbageDetector, a method to automatically analyze and classify short-answer responses as being valid or garbage. We show that GarbageDetector achieves excellent classification accuracy on real-world student data. Using data from several high school AP Biology and Physics classes, we present evidence that that providing valid short-answer responses creates a positive educational benefit on later practice.

Keywords

Best educational practices, Cognitive psychology, Machine learning, Natural language processing, Mixed effect modeling

1. INTRODUCTION

1.1 Overview and Motivation

Within education, it is critical that students not only acquire new knowledge, but that they are able to use that knowledge. Thus, an important part of the learning process is practicing the use of learned information by recalling that information from memory. This process, referred to as retrieval practice in the cognitive psychology literature [11], has been demonstrated to be a powerful and efficient way to improve student learning and retention [12].

Implementing retrieval practice within educational technology is straightforward – simply have students answer questions about the target material. The most commonly used question format is multiple-choice, primarily because multiple-choice responses are easy to machine score. While multiple-choice may be a great option for assessment, it is worth asking whether is is the best option for improving learning. Indeed, multiple-choice questions are oft-criticized because they are perceived to require only shallow recognition processes to complete.

An alternate format to multiple choice is short-answer. Short answer questions are not as frequently used as multiple-choice because natural language responses are difficult to machine score. However, they afford difficult reconstructive cognitive processes that are believed to be beneficial for student learning. Thus, it is often expected that short-answer responses are more beneficial to student learning. Interestingly, experiments examining the relative benefits of short-answer and multiple-choice questions on learning are mixed. Often, there is little to no difference between the two formats [22], but short-answer questions appear to be more beneficial than multiple-choice in scenarios where correct answer feedback is provided [10, 14]. Thus, short-answer may afford better learning for retrieval, but better understanding of the nuances of short-answer questions is required.

One factor that has not been examined in prior research is how the quality of short-answer responses provided by students contribute to learning. In online educational settings where students lack oversight, students do not always take the time to craft thoughtful short-answer responses. Instead, they often opt to to quickly enter a "garbage" response to advance their progress or view feedback. Indeed, surveys of college students found that when they practice retrieval on their own, they often avoid overt production of responses [26]. In this project, we examine how production of valid short-answer responses during retrieval practice influences learning.

A reasonable hypothesis is that students will derive greater learning benefits when they produce valid short-answer responses than when they do not. This hypothesis rests the assumption that producing garbage responses is an indication that the student was not actively engaged in retrieval processes. As mentioned previously, short-answer questions appear to be most effective when students are provided with subsequent feedback [10], suggesting that producing a short-answer response improves the processing of subsequent correct answer feedback. Moreover, several studies have found that engaging in retrieval improves subsequent encoding, even when the information recalled was not correct [12, 13]. This last part is important, because it suggests students need not enter a correct response to receive the benefits of short-answer questions. Indeed, Table 1 presents several real-world student responses to a biology question, that highlight the distinction between validity and correctness of these responses. In sum, we expect that producing valid short-answer responses will produce better learning, regardless of correctness.

We test the hypothesis that producing valid short-answer responses

Table 1: An example question taken from AP Biology with a set of actual responses provided by students. The examples cover both valid and garbage responses, with the valid responses including both correct and incorrect responses.

What is true about the energy released by the hydrolosis of ATP?		
Student Response	**Valid?**	**Correct?**
It powers many chemical reactions	Yes	Yes
It's short term	Yes	Yes
It's very high energy	Yes	No
It produces energy and water	Yes	No
A lot, surge, hyper	No	No
nope	No	No
asdlkfjas	No	No

improves learning with a large set of real user data, obtained from a web-based learning platform, OpenStax Tutor. In order to test this hypothesis, it is necessary to first determine whether the short-answer responses entered by students are valid. While this problem is easy at a small scale, it becomes much harder at a large scale.

Our contributions to this area of research are two-fold. First, we develop an automated method for classifying short-answer responses as being valid or garbage. Our method, which we dub *GarbageDetector*, relies on natural language processing (NLP) techniques to capture the salient information in short-answer responses and then to classify the response as being valid or garbage using supervised machine learning techniques. We show that this method performs extremely well on real-world student data, achieving a classification accuracy above 92%. Second, we further use GarbageDetector to automatically classify over 100,000 short-answer responses to questions collected during a pilot study across several high school science classrooms. Finally, we present evidence that crafting valid short-answer response to questions provide a strong learning benefit which translates to improved performance on later retrieval practice questions on the same topic. Our results demonstrate that this effect extends above and beyond the student's initial success in correctly answering multiple-choice questions.

1.2 Related Work

There are a variety of works that utilize extracted textual features from students' responses to predict future performance. Using higher-order language features to predict essay quality is discussed in [6]. The work in [18] analyzes students' textual responses and found that verbosity is an important predictor of their performance on the task at hand. The work described in [3] analyzes text of students' self-explanations and found that word occurrence statistics cannot be used to predict their responses to questions on its own. Neural networks are used to analyze student comments in [15] to predict their course grades. Student interactions in discussion forums of a massive open online course (MOOC) were analyzed in [24] where features such as topic composition of their posts were used to predict their performance on post-tests. Student dialogues with an automated tutoring system were used to estimate student prior knowledge in [23]. Our work differs significantly from these prior works in that we focus on using textual features to measure *future* learning outcome, i.e., long-term knowledge retention, as opposed to predicting *immediate* learning outcome or understanding levels. Moreover, we study the specific feature of response *validity* as a proxy for students' amount of retrieval effort, and its impact on learning.

The work in [25] analyzes student discussion forum posts in a MOOC and also trained a classifier to classify whether a post is on-task, and found that the quantity of on-task posts is a significant predictor of student learning gains. While related, our work is fundamentally different as we focus on analyzing the impact of the retrieval effort students put crafting valid responses to questions, rather than discussion dialogues, on future retrieval performance. Our task is significantly more complicated as the size of a student short-answer responses is often significantly shorter than a discussion forum post, making the final classification task more challenging.

2. VALIDITY CLASSIFICATION VIA GARBAGEDETECTOR

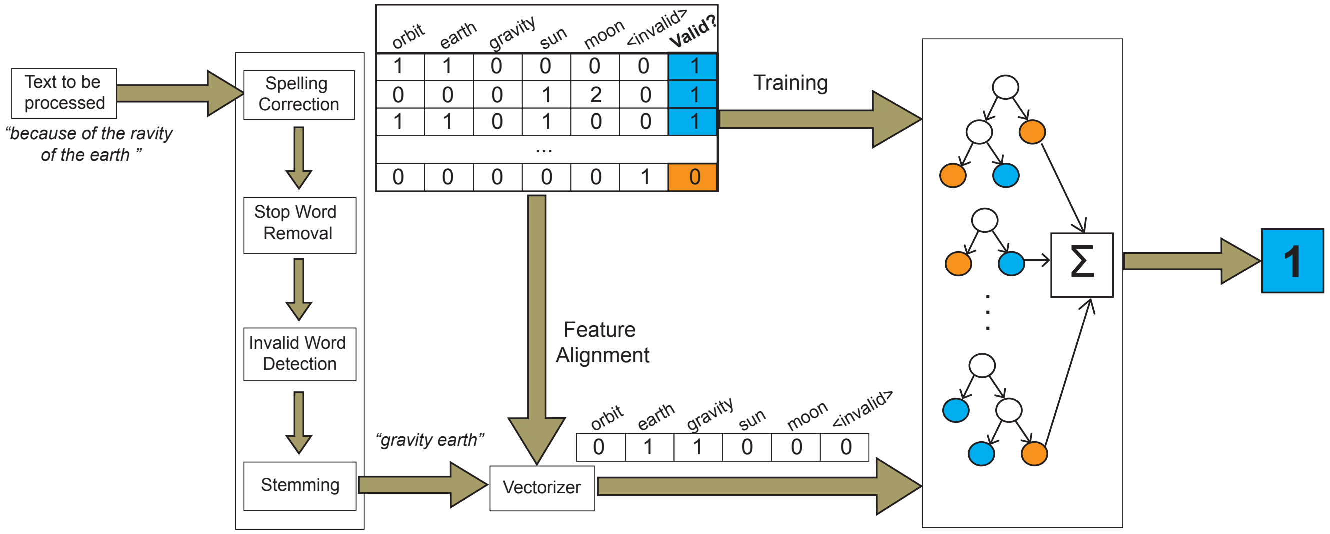
In this section, we describe GarbageDetector, our method for classifying short-answer responses as either valid or garbage. GarbageDetector first parses a student's short-answer response to extract and retain salient information. It then classifies the response using a binary classifier, which is trained on prior data consisting of a set of responses and ground-truth labels provided by human expert graders. A full block diagram of GarbageDetector is shown in Figure 1.

2.1 Parsing

Automatic validity classification of student short-answer responses is difficult for two primary reasons. First, as the name implies, these responses are extremely short (generally less than 10 words). Second, the vocabulary for a given student may be quite distinct from others. This results in a very sparse, high dimensional feature space that ultimately leads to poor classification accuracy. To overcome this challenge, we need to construct a parser that reduces the size of this feature space while retaining the information needed for classification.

To accomplish this feature space reduction, we first create a dictionary of acceptable words that are appropriate to use in responses, collected from both the general english language corpora as well as the domain-specific textbooks used in a given course. We then use this dictionary to automatically correct misspelled words, which are very common in student responses. While a human is able to identify these misspellings easily, it is more difficult for a machine to recognize that a word is misspelled or if it is simply a nonsensical word that does not appear in the dictionary. To combat this, we train a domain-specific spelling corrector similar to the work in [8] using our dictionary. This method combines the prior probability of seeing a given word as well as the edit distance [17] of the observed word from the set of all words in our dictionary. GarbageDetector does not correct to a word with an edit distance larger than 2 from every word in the dictionary to avoid false correction of nonsensical words, such as random character strings.

After spelling correction, the parser then removes any stop words (e.g., "the", "of", "is", etc.) from the response, since these words carry little semantic information regarding the validity of the response and greatly increase the size of the overall feature space. Following stop word removal, all words not found in our vocabulary dictionary are mapped together and replaced with a special label denoting that the words are unknown. As a final measure to reduce the feature space, the parser stems all recognized words using the Snowball stemmer [20]. As a concrete example, the words "biology" and "biological" are both reduced to the common stem "biolog".



orbit	earth	gravity	sun	moon	<invalid>	Valid?
1	1	0	0	0	0	1
0	0	0	1	2	0	1
1	1	0	1	0	0	1
...						
0	0	0	0	0	1	0

orbit	earth	gravity	sun	moon	<invalid>
0	1	1	0	0	0

Figure 1: A block diagram of GarbageDetector. In this example, the students initial response to a question about why the moon doesn't revolve in a circle around the sun is “because of the ravity [sic] of the earth” which, after parsing and classification, is determined to be valid.

2.2 Feature Construction

After the response is parsed, GarbageDetector converts it into a numerical feature vector using a simple bag of words (BOW) model [4], where each entry denotes the number of times a particular word is present in a response. Our method includes all 1-grams and 2-gram counts.

2.3 Classification

GarbageDetector employs a random forest classifier [9] trained on prior student responses. A random forest consists of a family of decision trees [21]. Each decision tree is trained on a subset of the data and learns a series of rules that help distinguish between valid and garbage short-answer responses. After each decision tree has classified the response, the results are aggregated and the most common label proposed by the individual decision trees is chosen for the final classification label. The overall accuracy of the random forest classifier is generally much higher than the accuracy of any one single decision tree. While training the random forest, we further make use of variable selection techniques [5] to reduce the feature space by selecting the most predictive features for classification.

2.4 Validation and Discussion

To validate the performance of GarbageDetector we employed a simple dataset consisting of student responses to questions in high school advanced placement (AP) Biology and standard (non-AP) high school Physics. This dataset consisted of over 20,000 short-answer responses that were manually labeled as valid or garbage by subject matter experts. We used leave-one-out cross validation to assess performance and trained aggregate all training examples separately at the chapter level. Our results are displayed in Figure 2, where we see an overall classification accuracy of 92%. We further quantified the importance of each step of the parser, by repeating the experiment without each individual step. We found that leaving out each any one step in the parser significantly reduced the overall classification accuracy. This results justifies the use of our full parser. Finally, we compared our classifier against a method that used our full parsing scheme but replaced the random forest classifier with a classifier that simply calculates the cosine similarity between the feature vectors of parsed student response and the text of the corresponding chapter. We then label a response as valid or garbage by comparing it against a threshold. We manually tuned this threshold to optimize performance and yet still experienced a 10% accuracy loss compared to the full GarbageDetector. This result justifies the use of the supervised classification method.

3. BENEFITS OF ENTERING VALID SHORT-ANSWER RESPONSES

We now turn our attention to evaluating the impact of providing valid short-answer responses on future learning outcomes using real-world educational data.

3.1 Dataset details

Our dataset is taken from the pilot run of our online learning platform, OpenStax Tutor [19], which was conducted during the 2015–2016 academic year. Tutor enables instructors to create retrieval practice assignments for their students as their classes progress. It has two important features relevant to the context of this paper. First, the questions use a hybrid answering format, consisting of first prompting the student to enter a short-answer response to the question and then, after the short-answer response has been submitted, prompting the student to select the correct answer to the question from a list of multiple-choice options. This feature provides students the opportunity to craft either a valid or garbage response to the question. The second important feature of our learning platform is that it automatically selects questions from previous assignments for spaced practice (see Figure 3 for an example). Briefly, spaced practice refers to spreading out learning over time, which has been known to improve long-term retention [7]. On any given assignment, in additional to the questions selected by the instructor, our online platform automatically presents questions from topics introduced on previous assignments. The purpose of this feature is to ultimately improve long-term knowledge retention, but we leverage these spaced practice observations as an opportunity to observe

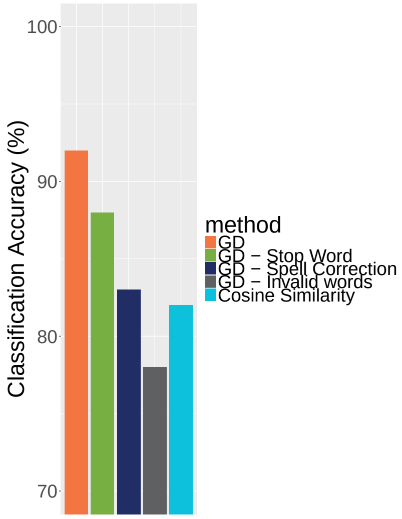


Figure 2: Comparison of classification accuracy for GarbageDetector (GD) along with parsing schemes that removed certain key features (Stop Word Removal, Spelling Correction, etc). We additionally compared against an unsupervised method that simply examined cosine similarity between the student response and the corresponding section in the textbook, tuned for optimal performance. GarbageDetector achieved the best performance against all of the other methods.

the effects of entering valid short-answer responses during earlier assignments on later spaced practice.

Our pilot study administered courses in AP Biology and standard (non-AP) Physics at 7 high schools in two separate school districts. A total of 207 students (74 AP Biology, 154 Physics) and 8 instructors (4 AP Biology, 4 Physics) participated in the pilot. While we do not have individual statistics on the students, aggregate statistics at the school level show that 85% of students are from minority populations, with 50% of students considered at-risk, and 60% considered economically underprivileged. AP Biology students were predominantly high school seniors, while Physics students were predominantly high school juniors. There are a total of 10,972 practice questions selected by instructors and a total of 1,644 spaced practice questions selected by Tutor in the dataset. All data was collected in accordance with the American Psychological Associations's Ethics Code.

3.2 Training and Classification

Since the number of responses is too large to be effectively labeled by humans, we automatically classify each short-answer response as either valid or garbage using the GarbageDetector method developed in Section 2. Our training data was obtained from a team of 7 subject matter experts – 6 on AP Biology and 1 on Physics who each labeled a subset of the responses as either valid or garbage. For AP Biology, 30% of the short-answer responses were labeled, for Physics only 5% of the short-answer responses were labeled. We train a separate classifier for each chapter of each textbook.

3.3 Analysis and Results

The setup of our analysis is visualized in Figure 3. Each section of an OpenStax textbook has a bank of associated practice questions, and each retrieval practice assignment in our dataset consists of some subset of textbook sections and their associated questions. We denote the initial questions when a student first does retrieval practice from a particular section as their *core* questions and use the term *spaced* question to denote the first time that a student encounters a question from a previously introduced textbook section during spaced practice. We hypothesize that entering valid short-answer responses during the initial core questions will improve performance on the topic when it is presented as spaced practice on a future assignment. We note that due to some degree of randomization different students in a class may receive spaced practice questions from different textbook sections, and there is no guarantee that each student will encounter a spaced question for every textbook section. In total, we have 1987 student-section observations for AP Biology and 4000 student-section observations for Physics. The median time between the last core question and the first spaced question for each topic is roughly 3 weeks.

We adopt a mixed effect logistic regression model [2, 16] for our analysis. This model uses two different sets of variables, termed random effects and fixed effects, to model a binary outcome variable. In our case, the binary outcome corresponds to whether the student selected the correct multiple-choice option on their spaced practice question for a given topic. The random effects correspond to nuisance quantities that are specific to each entity involved in the model, i.e., the individual student and textbook section effects. The random effects are modeled as simple intercept terms. The fixed effects correspond to the parameters of interest and, in our case, correspond to the number of correct multiple-choice responses and the number of valid responses on the initial core questions. The fixed effects are modeled as slope terms.

In summary, our model can be expressed mathematically as:

$$P(Y_{i,s}=1)=\Phi\left(\sum_j \alpha_j f_j^i+\sum_k r_k^i\right),$$

where $Y_{i,s} \in \{0,1\}$ denotes the binary-valued graded response of student i to the first spaced practice question from section s (with 1 denoting a correct response), f_j^i denotes the j^{th} fixed effect for student i, α_j denotes the slope term of the j^{th} fixed effect, r_k^i denotes the intercept term of the k^{th} random effect for student i, and $\Phi(x)=\frac{1}{1+e^{-x}}$ denotes the inverse logit link function. Concretely, we will consider 4 models:

- $\mathscr{M}_1$ considers only the random effects. This is an effective control model that we can use as a baseline for comparison.
- $\mathscr{M}_2$ considers the random effects and the number of correct multiple-choice responses that the student provided on their initial core questions for the given topic as a fixed effect.
- $\mathscr{M}_3$ considers the random effects and the number of valid responses that the student provided on their initial core questions for the given topic as a fixed effect.
- $\mathscr{M}_4$ considers the random effects and both the number of valid responses and number of correct multiple-choice responses that the student provided on their initial core questions as fixed effects.

We fit all four models to the AP Biology and Physics datasets separately. The results for AP Biology and Physics are shown on Table 2 and Table 3, respectively. In order to determine which

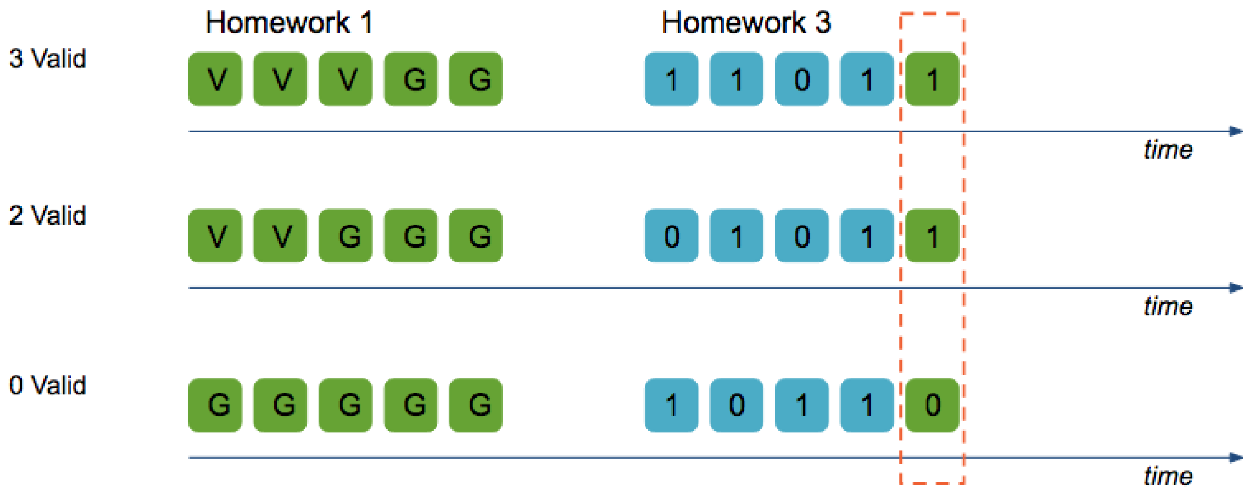


Figure 3: Overview of our analysis. Students respond to questions (boxes) across a series of topics (colors). On their first exposure to a topic, students provide either valid (V) or garbage (G) short-answer responses to the questions. Later, the students are presented with a spaced practice question drawn from a previous topic (dashed box) and provide either a correct (1) or incorrect (0) multiple-choice selection.

model provided the best fit, we used the Akaike information criterion (AIC) metric [1], which imposes a penalty that penalizes modes with too many parameters to prevent overfitting. Models with lower AIC values are deemed better than models with higher AIC values.

For AP Biology, $\mathscr{M}_3$ provided a reduction in AIC compared to $\mathscr{M}_2$, implying that the number of valid responses provided a better predictor of success than the number of correct multiple-choice selections. The coefficient for the number of valid responses is positive and statistically significant, which matches our hypothesis that more valid responses improves student retention. Moreover, $\mathscr{M}_4$ provides a reduction in AIC over $\mathscr{M}_2$ but not $\mathscr{M}_3$. This result implies that adding valid responses improves the model fit compared to using number of correct responses alone. We note however that for $\mathscr{M}_4$ the coefficient for the number of correct responses is essentially 0, again implying that $\mathscr{M}_3$ is the best model for this subject domain, i.e., the number of valid responses is a better predictor than the number of correct multiple-choice selections.

For Physics, we note that $\mathscr{M}_3$ does not reduce the AIC over $\mathscr{M}_2$, meaning that the number of correct responses alone does provide higher predictive power than the number of valid responses alone. However, the AIC of $\mathscr{M}_4$ is less than $\mathscr{M}_2$, and both coefficients are positive and statistically significant. This result implies that both factors together produce better modeling fitting.

Finally, to illustrate the effect of producing a valid free-form response on learning, we produced a visualization using the best models from AP Biology and Physics ($\mathscr{M}_3$ and $\mathscr{M}_4$, respectively). To produce this visualization, we took each student in the dataset and set their number of valid responses for each topic to some constant value. We then used our model to predict whether each student would have answered their spaced practice problem correctly as a function of the number of valid responses as well as the student-specific random effects. We repeated this procedure over a reasonable range corresponding to the actual number of exercises on assignments. The resulting visualization is shown on Figure 4, which shows a significant difference between a hypothetical student who makes no effort to provide valid responses during their core retrieval exercise and those who provide valid responses to all questions. Concretely, students who who provide all valid responses are predicted to have a 20% improvement to their chance of answering their spaced practice exercise correctly.

Table 2: Summary of AP Biology Data Models

	Dependent variable:			
	Correct on Spaced Practice			
	(1)	(2)	(3)	(4)
Number Core Correct		0.030^{*} (0.016)		-0.009 (0.027)
Number Core Valid			0.034^{**} (0.013)	0.040^{*} (0.023)
Constant	0.613^{***} (0.075)	0.467^{***} (0.107)	0.427^{***} (0.105)	0.437^{***} (0.109)
Observations	1,987	1,987	1,987	1,987
Log Likelihood	$-1{,}278.010$	$-1{,}276.102$	$-1{,}274.653$	$-1{,}274.599$
Akaike Inf. Crit.	2,562.019	2,560.203	2,557.305	2,559.199

Note: *p<0.1; **p<0.05; ***p<0.01

4. CONCLUSIONS

We have developed GarbageDetector for classifying student open-form responses to questions as being either valid (on-topic) or garbage (off-topic) using a combination of intelligent parsing and supervised classification. We have shown that this method works well and can accurately classify student short-answer responses across two separate subject domains.

We have also presented evidence that students who spend time crafting thoughtful responses show improved learning outcomes, measured by performance on later spaced repetition of the same topic.

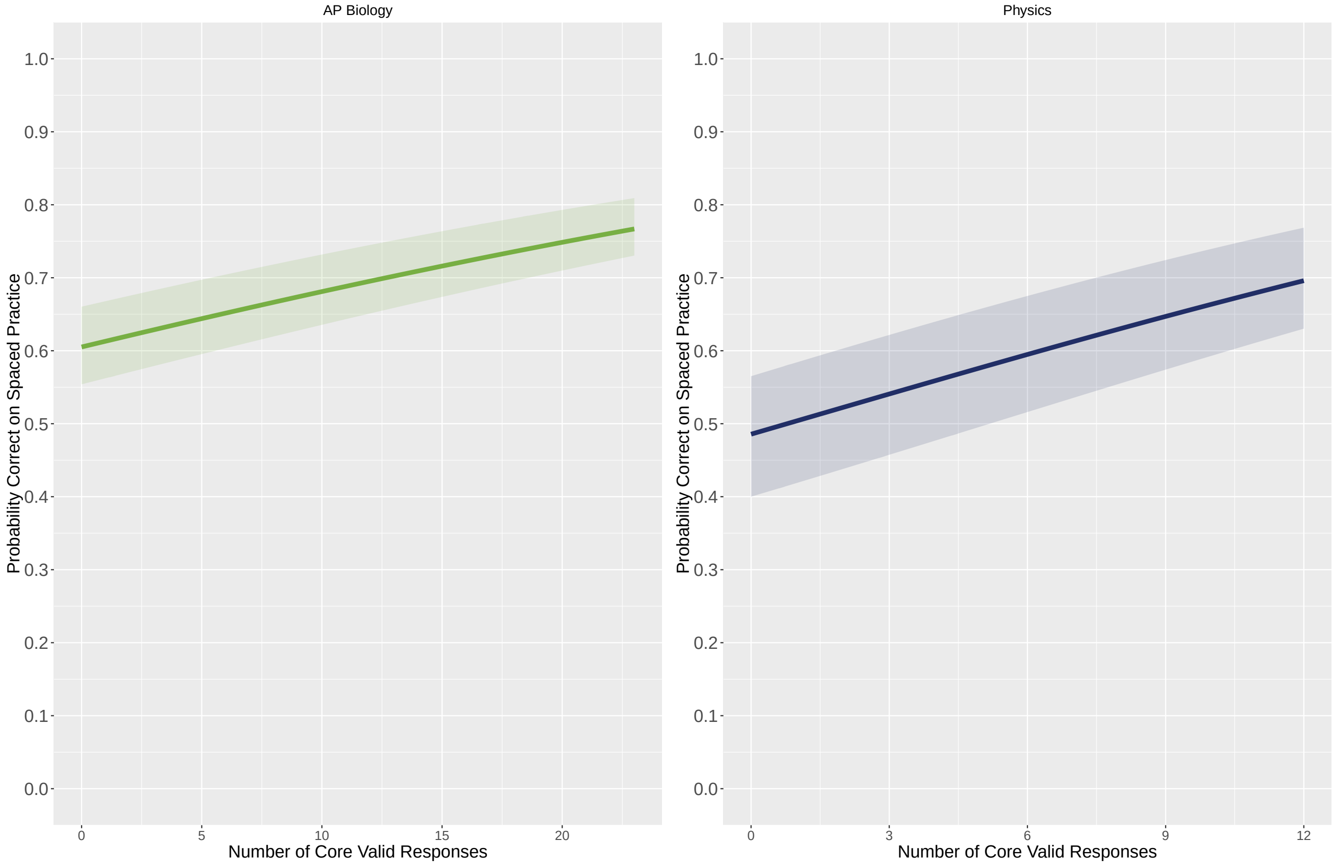


Figure 4: Predictive modeling showing the relationship between the number of valid short-answer responses and the probability of success on later spaced practice exercises. The bold line shows the average across all students, the shaded region shows confidence intervals. GarbageDetector predicts grade differences of up to 20% between students who never enter in valid responses and those who always do.

Table 3: Summary of Physics Data Models

	Dependent variable:			
	Correct on Spaced Practice			
	(1)	(2)	(3)	(4)
Number Core Correct		0.082***		0.076***
		(0.013)		(0.013)
Number Core Valid			0.097***	0.078***
			(0.023)	(0.022)
Constant	0.002	$-$0.316***	$-$0.105	$-$0.377***
	(0.074)	(0.087)	(0.079)	(0.089)
Observations	4,000	4,000	4,000	4,000
Log Likelihood	$-$2,703.761	$-$2,682.312	$-$2,693.697	$-$2,675.836
Akaike Inf. Crit.	5,413.522	5,372.623	5,395.394	5,361.672
Note:				*p<0.1; **p<0.05; ***p<0.01

The results that we have derived in this work are the result of searching for patterns in existing data and relied on students deciding of their own volition whether or not to enter a valid short-answer response. Future research in this area will involve more highly controlled study in which the opportunity to enter a short-answer response will be controlled by our learning system. This will allow us to create two test cohorts of students, namely those who had the opportunity to enter a short-answer response and those that did not. This will allow us greater control over our experimental setup and aid in the interpretation of our final result.

5. ACKNOWLEDGMENTS

Thanks to the Art Ciocca, the Laura and John Arnold foundation, and John and Ann Doerr for supporting this research. Thanks also to Micaela McGlone, Debshila Basu Malik, and Alicia Chang for their help in conducting the pilot studies, preparing the data, performance analysis, as well as many helpful discussions regarding this work.

6. REFERENCES

[1] H. Akaike. Likelihood of a model and information criteria. *J. Econometrics*, 16(1):3–14, 1981.

[2] D. Bates, M. Maechler, B. Bolker, S. Walker, et al. LME4: Linear mixed-effects models using Eigen and S4. *R Package*, 1(7), 2014.

[3] S. Bhatnagar, M. Desmarais, N. Lasry, and E. S. Charles. Text classification of student self-explanations in college physics questions. In *Proc. 9th Intl. Conf. on Educational Data Mining*, pages 571–572, July 2016.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Drichlet allocation. *J. Machine Learning Research*, 3:993–1022, Jan. 2003.

[5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] S. Crossley, K. Kyle, D. S. McNamara, and L. Allen. The importance of grammar and mechanics in writing assessment and instruction: Evidence from data mining. In *Proc. 7th Intl. Conf. on Educational Data Mining*, pages 300–303, July 2014.

[7] F. N. Dempster. Spacing effects and their implications for theory and practice. *Educational Psychology Review*, 1:309–330, 1989.

[8] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009.

[9] T. K. Ho. Random decision forests. In *Proc. 3rd Intl. Conf. Document Analysis and Recognition*, volume 1, pages 278–282. IEEE, 1995.

[10] S. Kang, K. McDermott, and H. Roediger. Test format and corrective feedback modify the effects of testing on long-term retention. *European J. Cognitive Psychology*, 19:528–558, 2007.

[11] J. Karpicke and J. Blunt. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331:772–775, 2011.

[12] J. Karpicke and P. Grimaldi. Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review*, 24:401–418, 2012.

[13] N. Kornell. Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1):106, 2014.

[14] J. L. Little, E. L. Bjork, R. A. Bjork, and G. Angello. Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, 23:1337–1344, 2012.

[15] J. Luo, E. Sorour, K. Goda, and T. Mine. Predicting student grade based on free-style comments using Word2Vec and ANN by considering prediction results obtained in consecutive lessons. In *Proc. 8th Intl. Conf. on Educational Data Mining*, pages 396–399, June 2015.

[16] C. E. McCulloch and J. M. Neuhaus. *Generalized Linear Mixed Models*. Wiley Online Library, 2001.

[17] G. Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.

[18] B. D. Nye, M. Hajeer, C. Forsyth, B. Samei, X. Hu, and K. Millis. Exploring real-time student models based on natural-language tutoring sessions: A look at the relative importance of predictors. In *Proc. 7th Intl. Conf. on Educational Data Mining*, pages 253–256, July 2014.

[19] OpenStaxTutor. https://openstaxtutor.org/, 2017.

[20] M. F. Porter. Snowball: A language for stemming algorithms, 2001.

[21] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[22] M. Smith and J. Karpicke. Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, 22:784–802, Sep. 2013.

[23] D. Stefanescu, V. Rus, and A. C. Graesser. Towards assessing students' prior knowledge from tutorial dialogues. In *Proc. 7th Intl. Conf. on Educational Data Mining*, pages 197–200, July 2014.

[24] S. Tomkins, A. Ramesh, and L. Getoor. Predicting post-test performance from online student behavior: A high school MOOC case study. In *Proc. Intl. Conf. Educ. Data Min.*, pages 239–246, June 2016.

[25] X. Wang, D. Yang, M. Wen, K. Koedinger, and C. P. Rosé. Investigating how student's cognitive behavior in mooc discussion forums affect learning gains. In *Proc. 8th Intl. Conf. on Educational Data Mining*, pages 226–233, June 2015.

[26] K. T. Wissman, K. A. Rawson, and M. A. Pyc. How and when do students use flashcards? *Memory*, 20(6):568–579, 2012.