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

Andrew Waters, Phillip Grimaldi, Andrew Lan, and Richard Baraniuk

Abstract—Many studies in cognitive psychology have shown stronger learning benefits from short-answer questions than multiple-choice questions in student practices. However, multiple-choice is still, by far, the most popular form of assessment questions due to ease of grading. A hybrid method, in which students must first provide a short-answer response and then select from a list of multiple-choice options, offers a nice compromise that would allow students the learning benefit of short-answer responses while still keeping the grading simple and scalable. However, students often bypass the short-answer by entering invalid (off-topic) responses to more quickly access the final multiple-choice selection, thereby bypassing any learning benefits offered by short-answer responses. In this work, we develop a novel method, GarbageSniffer, to automatically analyze and classify short-answer responses to questions as being valid (on-topic) or invalid (off-topic). We show that our method achieves excellent classification accuracy using real-world student data. We then analyze the impact of entering valid short-answer responses on future learning outcome using a separate real-world educational dataset taken from several high school AP Biology and Physics classes. In particular, we present evidence that students who make an effort to provide valid short-answer responses derive an educational benefit on later practice that is not seen in students who enter invalid short-answer responses.

Keywords—Natural language processing, Machine learning, Cognitive psychology, Best educational practices.

I. INTRODUCTION

A. Overview and Motivation

A core part of the student learning experience is practicing the application of learned knowledge, often in the form of homework or quizzes. This experience, referred to as retrieval practice [11] in the cognitive psychology literature, has been demonstrated to be a powerful and efficient way to improve student learning and retention [12].

Most methods for implementing retrieval practice fall into one of two categories: multiple-choice questions and short-answer questions. A reasonable question to ask is which one of these methods conveys a stronger learning benefit to students. Retrieval effort is thought to be an important factor governing the effectiveness of retrieval, with increased effort resulting in improved learning outcomes [22]. Because of this, one can intuitively see why short-answer questions would likely outperform multiple-choice questions in terms of learning. Short-answer questions generally require more effort to solve given that a student must reconstruct knowledge from memory with limited retrieval cues. Conversely, when answering multiple-choice questions, students need only recognize the answer from a list of available options.

Despite the benefits of short-answer questions, they are less used than multiple-choice questions simply because of the effort required to grade them. Human instructors/ graders cannot reasonably grade all student responses well, especially as class sizes become large. Further, methods for automatically grading short-answer responses by machine have not achieved the level of accuracy needed for widespread applications [13] or are limited to grading longer essays [26].

An attractive format for retrieval practice is a hybrid scheme [20], in which students first provide a short-answer response and then select their final response from a multiple-choice list. The hybrid format seems to be a good compromise, given the fact that both short-answer and multiple-choice formats seem to produce positive benefits for learning while ensuring that grading is not cumbersome. Studies on the benefits of the hybrid format are sparse, but the hybrid format has been shown to enhance learning relative to short-answer or multiple-choice. A major issue with the hybrid format, however, is that in the real world students tend to not take the short-answer portion seriously.

At OpenStax [19], we have collected data from thousands of students using a hybrid format for retrieval practice. A disturbingly common trend is that students will enter invalid responses to simply get past the short-answer portion and view the multiple-choice selections. The reasons why students do this are beyond the scope of this paper. The important thing to consider is the end result: entering invalid responses bypasses the retrieval processes that the short-answer portion is intended to encourage.

A solution to the problem of students bypassing the short-answer portion of the hybrid format would be to check the short-answer response and determine if it is valid. In this work we use the term valid to denote that a response is on-topic and related to problem and subject material, rather than referring to whether or not the response is correct. Table I presents several examples that highlight the distinction between validity and correctness.

B. Contributions

In this work, we derive an automated method for classifying short-answer responses as being valid or invalid. Our method,
which we dub GarbageSniffer, relies on Natural Language Processing (NLP) techniques to preserve salient information in short-answer responses and then classify the response using supervised machine learning techniques. We show that this method performs extremely well on real-world student data, achieving a classification error of around 5%.

We further use GarbageSniffer to automatically classify tens of thousands of short-answer responses to questions collected by OpenStax Tutor during a pilot study. We then present evidence that students who took the time to thoughtfully craft valid short-answer responses received a very strong learning benefit that translates into later performance on questions drawn from the same subject material. This result, we show, extends above and beyond the student’s initial success in correctly selecting the correct multiple-choice option on their initial practice.

II. RELATED WORK

NLP has been widely applied and has enjoyed great success in a wide range of applications including information retrieval, collaborative filtering, etc [15]. Early developments include term frequency-inverse document frequency (TF-IDF) [24] and the bag-of-words (BOW) model (see, e.g., [3]). The BOW model, in particular, models the word counts in textual documents and has enjoyed great success in tasks like document indexing and compression. Recent developments include the continuous bag-of-words (CBOW) model, which uses recurrent neural networks [17] to embed words into a low-dimensional vector space. This model excels at preserving semantics and other linguistic features of the original text.

NLP has also enjoyed great success in the learning sciences. The work in [5] first suggests to analyze textual data using NLP techniques and a great deal of work has followed. The works in [25], [27] use these techniques to analyze student dialogues and posts in discussion forums, while other works (e.g., [6]) extract higher-order language features from student essays and use them to predict essay qualities.

A number of works focus on the problem of predicting student performance using textual features extracted from various sources. Specifically, the work in [25] extracts features from student dialogues to predict student grades. The work in [18] analyzes textual student responses to questions and found that verbosity is an important predictor of student grades. The work in [14] analyzes short student comments and uses a neural network to predict student grades. This work significantly differs from prior art in that we are focused on using textual features to measure future learning outcomes as opposed to predicting current grades. Moreover, we study the specific feature of response validity and its impact on learning.

III. VALIDITY CLASSIFICATION VIA GARBAGESNIFFER

In this section, we describe GarbageSniffer, our method for classifying valid and invalid short-answer responses. GarbageSniffer first parses a student’s short-answer response to extract and retain salient information. GarbageSniffer then classifies the response using supervised learning techniques. A full block diagram of GarbageSniffer is shown in Figure 1.

A. Parsing

Let \( r = [w_1, w_2, \ldots, w_K] \) denote a student’s response to a given problem with \( w_k, k \in 1, \ldots, K \), denoting the \( k^{th} \) word in the response. Our goal is to develop a classifier \( C \) that takes as input a response \( r \) and classification \( C \in \{0,1\} \) with 1 denoting a valid response and 0 denoting an invalid response.

The potentially large number of unique words in student responses would inevitably result in a high-dimensional feature space that would lead to poor classification accuracy. To overcome this, we first intelligently parse out the student responses to reduce the overall feature space while still retaining the key information in the responses needed to classify valid/invalid responses. Concretely, the parser generates a parsed vector \( \hat{r} = [\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_L] \) with \( w_k, k \in 1, \ldots, L \leq K \), denoting the the \( \ell^{th} \) parsed term. The parser operates on each word individually and performs the following operations:

- **Stop word removal:** In this step, the parser remove typical stop words (e.g., “the”, “of”, “is”, etc.) from the responses. These words carry little valuable information and, if left in the response, considerably enlarge the feature space. Each word in the response is checked against a pre-defined list of stop words and is removed if it is found in the list.

- **Numeric Tagging:** Many subjects, such as physics, require students to use mathematical and numerical reasoning to solve problems. Since we are only interested in whether a student’s response is valid and not whether the student’s response is correct, it is useful to reduce the feature space of mathematical expressions and numbers. To do this, we classify such terms into a small number of classes. Mathematical expressions (e.g., “\( y = e \land x \)” are simply tagged as being mathematical expressions. Numbers are tagged according to their type and include integers, decimals, complex numbers, and Roman numerals.

- **Spelling correction:** One of the largest impediments to detecting valid responses is the high frequency of misspelled words in student short-answer 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 just an entirely new word that it has not seen previously. To improve this, we train a domain specific spelling corrector similar to the work in [8] using a training corpus consisting of the both general text as

---

**TABLE I:** An example question taken from OpenStax AP Biology with a set of actual responses provided by students. The examples cover both valid and invalid responses, with the valid responses including both correct and incorrect responses.

<table>
<thead>
<tr>
<th>Question: What is true about the energy released by the hydrolysis of ATP?</th>
<th>Valid?</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td>It powers many chemical reactions</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>It’s short term</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>It’s very high energy</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>It produces energy and water</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>A lot, surge, hyper</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>nope</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>audifjjas</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
well as the target textbook for the subject. This method combines the prior probability of seeing a given word as well as the edit distance of the observed word from the set of all words in our lexicon. GarbageSniffer does not correct to a word with an edit distance larger than 2.

- **Invalid word tagging:** In this step, the parser identifies and tags invalid words. If spelling correction does not succeed and the word is not found in our dictionary then it will be tagged as invalid. This step is primarily meant to catch students who type random characters in their responses.

- **Stemming:** In this step, the parser further reduces the feature space by combining words with the same stem. As a final measure to reduce the feature space, the parser stems all valid words. As a concrete example, the words “biology” and “biological” are both reduced to the common stem “biolog”. In our approach, we make use of the Snowball stemming package [21].

**B. Feature Extraction and Vectorizing**

After a response is parsed, GarbageSniffer converts it into a numerical feature vector using a simple bag of words (BOW) model [3], where each entry denotes the number of times a particular word is present in a response. Our model includes all 1-grams, 2-gram, and 3-gram counts. We experimented with other feature models such as continuous bag of words (CBOW) [17] and found no improvement in classification accuracy.

**C. Classification and Training**

GarbageSniffer employs a random forest classifier [10], [9] trained on prior student responses to classify new short-answer responses. A random forest consists of a family of decision trees [23] that vote individually on the extracted feature vector.

The final output of a random forest classifier is a proportion \( p \in [0,1] \) denoting the number of trees that classified the response as valid. For simplicity, we classify a response as valid if \( p > 0.5 \). We use variable selection techniques [4] as a final measure to reduce the feature space to the most relevant features for classification.

An important design consideration is the scope of responses that will be used to train the valid response classifier. As most textbooks, OpenStax problems can be aggregated at the subject, chapter, section, or individual problem level. By aggregating questions and responses at a high level we can potentially aggregate a large number of training examples but could also potentially suffer due to extremely large variation in short-answer response features. Alternatively, creating a classifier for each individual question is an attractive option due to specificity but could result in too few training exemplars for the classifier to work well on new test cases. We analyzed the classification error for all four of these cases using a training dataset consisting of 15 high school participants (HSPs) enrolled in an OpenStax summer program who answering over 1000 questions in AP Biology and Physics. A team of human experts manually classified each response as either valid and invalid for training and validation purposes. We compute classification error using leave-one-out cross validation for each of the four aggregation cases and display our result in Figure 2. We see that training at the chapter level (followed closely by training at the section level) resulted in the best classification error of around 5\%, meaning that we can correctly classify 19 out of 20 short-answer responses correctly. We note that this is a large improvement over the state-of-the-art rate for grading short-answer responses [13].

**IV. EDUCATIONAL BENEFITS OF ENTERING VALID SHORT-ANSWER RESPONSES**

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

Our dataset is collected from a set of high school AP Biology and (non-AP) Physics courses piloted on OpenStax Tutor. OpenStax Tutor is a web-based learning platform that is companion to OpenStax online textbooks. OpenStax Tutor enables instructors to create retrieval practice assignments relevant to current class discussions. OpenStax Tutor has two important features relevant to the current discussion: 1) The questions use a hybrid format (see Figure 3) and 2) Tutor automatically presents questions from previous assignments for spaced practice. Briefly, spaced practice refers to the spreading of learning out over time, which has been known to improve long-term retention [7]. On any given assignment, OpenStax Tutor automatically presents questions from two assignments ago. The purpose for this feature is to improve ultimate learning of the spaced content, but we leveraged these spaced practice observations as an opportunity to observe the effects of entering valid free form responses during previous assignments. We hypothesize that producing valid short-answer responses during the initial core assignment will improve performance on the topic when it was presented on a future assignment for spaced practice. There are a total of 13,881 core practice problems with both a short-answer response and a multiple-choice selection and a total of 2,062 spaced practice problems in the dataset.

The pilot study was conducted in seven high school campuses across two separate school districts. There were four instructors participating in AP Biology as well as four instructors for Physics. Each instructor taught between one and five periods of a particular course. A total of 207 students participated in the study. While we do not have individual statistics on the students, aggregate statistics at the school level show that roughly 85% are from minority populations, nearly 50% are considered at-risk, and 60% are considered economically disadvantaged. AP Biology students were predominantly high school seniors, while Physics students were predominantly high school juniors. All data was collected in accordance with the American Psychological Association’s Ethics Code.

We are interested in analyzing whether or not students entering in valid short-answer responses during their initial core practice have an increased chance of success on the multiple-choice selection in the later spaced practices on the same topic. Since the number of responses is too large to be effectively labeled by humans, we automatically classify each core short-answer response as either valid or invalid using the GarbageSniffer method developed in Section III. We note that there is considerable variation in the data as far as valid/invalid responses and the student’s graded multiple-choice selection. We display summary statistics in Table II and note that there is not a high correlation between the validity of the short-answer response and the accuracy of the final multiple-choice selection.

B. Model Selection and Control Variables

We adopt a mixed effect logistic regression model [2], [16] to analyze the impact of short-answer responses on later spaced practice performance. 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 the correctness of students’ multiple-choice selection on each spaced practice problem. The random effects correspond to nuisance quantities that we cannot control, namely the student, class instructor, and topic effects. The random effects are modeled as simple intercept terms. Fixed effects are the parameters of interest and in our case correspond to the number of correct MC selections and the number of valid responses for the initial core problems. The fixed effects are modeled as slope terms. We also experimented with the number of homework problems attempted as an additional fixed effect but this provided no improvements in fitting the model to our data.

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 if student $i$ answered spaced practice problem $s$ incorrectly/correctly, $f_j^i$ denotes the $j^{th}$ fixed effect for student $i$, $\alpha_j$ 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.

It is very important to properly select which effects will be considered in the model to avoid under/over-fitting as well as to ensure that our results are meaningful. In particular, we must justify the inclusion of certain variables quantitatively. To do this, we will select models that minimize the Akaike Information Criterion (AIC) [1]. Concretely, we will consider 4 models:

<table>
<thead>
<tr>
<th>Correct MC</th>
<th>Valid Response</th>
<th>Invalid Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect MC</td>
<td>36%</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>13%</td>
<td>21%</td>
</tr>
</tbody>
</table>
$M_1$ considers only the random effects. This is an effective control model that we can use as a base model for comparison.

$M_2$ considers the random effects and the number of valid responses that the student provided on their initial core problems for the given topic.

$M_3$ considers the random effects and the number of correct multiple-choice selections that the student provided on their initial core problems for the given topic.

$M_4$ considers the random effects and both the number of valid responses that students provided on their initial core problems and the number of correct multiple-choice selections for the given topic.

We ran all of these models on the dataset and compiled the results in Table III for AP Biology and Table IV for Physics.

For AP Biology, the best model is $M_2$, in which only the number of valid responses that the student has submitted on their initial core problems.

For Physics the best performing model is $M_3$, which depends only on the number of correct multiple-choice selections submitted on the core problems. An analysis of the data shows that this effect is primarily due to poor data from one Physics instructor who did not use the course software correctly and did not emphasize the short-answer response feature. When we remove this instructor and re-run the model analysis we obtain the results in Table V, where we now see a very strong preference for $M_2$.

C. Results

Now that we have identified that the best model for predicting student success is the one in which we consider the number of core valid short-answer responses along with the random effects ($M_2$), we are ready to investigate the impact of this model on student learning and success. We do this via predictive modeling. Concretely, we take each student in the dataset and set their number of valid responses for each topic to some constant value. We then allow our model to predict if 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 repeat this procedure over a reasonable range corresponding to the actual number of homework problems on assignments. We display these results in Figure 4 for both AP Biology and Physics. We see that model predicts a significant difference in success between students who do not take the time to craft valid responses and those who predominantly do, causing potentially more than a full letter grade of improvement between the extreme cases.

V. CONCLUSION

We have developed GarbageSniffer for classifying student open-form responses to questions as being either valid (on-topic) or invalid (off-topic) using a combination of intelligent
Fig. 4: Predictive modeling for the number of valid short-answer responses. The bold line shows the average across all students, the shaded region shows quantiles. GarbageSniffer predicts grade differences of up to 20% between students who never enter valid responses and those who always do.

### TABLE V: Summary of Physics Data Models (filtered)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Core Valid</td>
<td>0.23***</td>
<td>0.23***</td>
<td>(0.08, 0.39)</td>
<td>(0.06, 0.40)</td>
</tr>
<tr>
<td>Number of Core Correct</td>
<td>0.09</td>
<td>0.01</td>
<td>(−0.08, 0.27)</td>
<td>(−0.17, 0.19)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.05</td>
<td>−0.29*</td>
<td>−0.15</td>
<td>−0.30</td>
</tr>
<tr>
<td></td>
<td>(−0.21, 0.31)</td>
<td>(−0.61, 0.03)</td>
<td>(−0.59, 0.29)</td>
<td>(−0.73, 0.12)</td>
</tr>
<tr>
<td>Observations</td>
<td>567</td>
<td>567</td>
<td>567</td>
<td>567</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>−386.56</td>
<td>−382.64</td>
<td>−385.98</td>
<td>−382.64</td>
</tr>
<tr>
<td>Akaike Inf. Crit.</td>
<td>779.11</td>
<td>773.28</td>
<td>775.97</td>
<td>775.27</td>
</tr>
</tbody>
</table>

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

We have shown that this method works well and can robustly classify student answers in two separate subject domains.

We have also presented strong evidence that students who spend time crafting thoughtful responses to questions receive an educational benefit on later spaced repetition of the same topic, even when the submitted answer is ultimately incorrect. We have shown that this benefit can be quite significant, potentially beyond a full letter grade of improvement, and explains student performance better than simply analyzing the number of correct responses submitted during initial practice.

Future work in this area will involve designing a more rigorous, controlled study to assess the impact of open-form responses. This would involve not allowing some students to use the short-answer response feature as well as controlling for the number of valid responses entered by the remaining students. It will further be insightful to see if we can build features into OpenStax Tutor that can foster this behavior in students, such as alerting a student or their instructor if we believe that they are not making the effort to think through their answer in a meaningful way.

### VI. ACKNOWLEDGMENTS

Thanks to the Laura and John Arnold foundation and Art Ciocca 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.

### REFERENCES


