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

Andrew Waters
OpenStax, Houston, TX
aew2@rice.edu

Phillip Grimaldi
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

Educational technology commonly leverages multiple-choice questions for student 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 often produce off-topic (or invalid) responses that are off-topic and do not relate to the question being answered. In this study, we consider the effect of entering on-topic short-answer response on student learning and retention. To do this, we first develop a machine learning method to automatically label student open-form responses as either valid or invalid using a small amount of hand-labeled training data. Then, using data from several high school AP Biology and Physics classes, we present evidence 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

An important part of the learning process is recalling learned information from memory [3]. In most educational situations, this practice is accomplished by asking students practice questions related to the learning material. In online learning, multiple-choice questions are by far the most common, following by short-answer questions. While multiple choice questions are attractive due to the ease of machine scoring, 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 [7]. Short-answer responses, by contrast, are generally believed to have a stronger learning benefit to students as they afford more difficult reconstructive cognitive processes.

Prior experiments examining the relative benefits of multiple-choice and short-answer have been mixed, with short-answer questions generally found to improve learning only when subsequent feedback is provided [2, 4]. One factor that has not been examined in prior research, however, 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 an off-topic response to advance their progress or view feedback.

We hypothesize that students derive greater learning benefits when they produce valid short-answer responses than when they do not, even when those valid responses are incorrect. While it is possible to hand-label student responses as valid or invalid for a small number, it is not feasible to do this at large scale. To circumvent this scalability issue, we devise a machine-learning based classifier trained on a small number of hand-labeled exemplars. We then leverage this classifier to analyze the impact of entering valid responses on learning.

2. AUTOMATIC VALIDITY CLASSIFICATION

Due to the large number of words in student responses, our method for automatically classifying student short-answer responses as valid or invalid begins with parsing to reduce the overall size of the feature space. First, we attempt simple spelling correction for each word of a student's response. Following spelling correction, which strip common stopwords (e.g. of, as, is, etc) and replace any nonsensical words (e.g., random keyboard presses) with a specially defined tag, which has the effect of mapping all unknown words to the same label. Finally, we stem acceptable words in a student responses to further reduce the dimensionality of our feature space. Finally, we convert the parsed student response to a numerical feature vector using a bag-of-words model.

Following parsing, we employ a random forest [1] to classify each student response as either valid or invalid. We measured the performance of our method using 5-fold cross-validation on $20,000$ hand-labeled responses and found our accuracy to be 95%.

3. ANALYSIS OF VALID RESPONSES ON LEARNING

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

Our dataset is taken from a pilot study of our online learning platform, OpenStax Tutor [6], which was conducted during the 2015–2016 academic year. OpenStax Tutor has two important features relevent to our discussion. First, it uses a hybrid answering format [7] that first requires students to enter a short-answer response to the question and requires the student to select the correct answer from a multiple-choice list. Second, OpenStax Tutor employs a concept known as spaced practice, which automatically assigns questions to students on material that they have learned in 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 the effects of entering valid short-answer responses on later practice.

The pilot consisted of two separate high school courses, AP Biology and standard (non-AP) Physics. A total of 207 students (74 AP Biology, 154 Physics) and 8 instructors (4 AP Biology, 4 Physics) participated in the pilot. There are roughly 100,000 short-answer responses on initial practice problems, and 20,000 of these answers were hand-labeled by subject matter experts as being valid or invalid responses to the given question. The average spaced practice problem occurs roughly 3 weeks after the initial practice on the topic is complete.

To analyze the impact of entering valid open-form responses we adopt a mixed effect logistic regression model [5]. Our binary outcome is whether or not the student answered the spaced practice question for a given topic correctly. Our random effects (R) are nuisance quantities for student ability, topic difficulty, and instructor quality. We examine two different fixed effects in our model: M, the number of multiple-choice questions that a student answered correctly on a given topic and V, the number of valid short-answer responses that a student provided on a given topic.

We consider four separate models for student success on spaced practice questions. Each model includes the random-effects R. We then separately consider the effects of the fixed effects M and V as well as considering both fixed effects jointly. We fit all four models to the AP Biology and Physics datasets separately. The results for AP Biology and Physics are shown on Table 1 and Table 2, respectively. In order to determine which model provided the best fit, we used the Akaike information criterion (AIC) metric, 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, we found that the $R+V$ model achieved the lowest AIC 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. For Physics, we note that $R+M+V$ provides the lowest AIC value, and is significantly better than considering $R+M$ alone. This implies that both factors together produce better modeling fitting.

Table 1: Summary of AP Biology Data Models

	Dependent variable:			
	Correct on Spaced Practice			
	(R)	$(R+M)$	$(R+V)$	$(R+M+V)$
Number Core Correct		0.030^{*}		-0.009
		(0.016)		(0.027)
Number Core Valid			0.034^{**}	0.040^{*}
			(0.013)	(0.023)
Constant	0.613^{***}	0.467^{***}	0.427^{***}	0.437^{***}
	(0.075)	(0.107)	(0.105)	(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

Table 2: Summary of Physics Data Models

	Dependent variable:			
	Correct on Spaced Practice			
	(R)	$(R+M)$	$(R+V)$	$(R+M+V)$
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

4. CONCLUSIONS

We have developed a machine-learning based method for classifying student open-form responses to questions as being either valid (on-topic) or invalid (off-topic) using a combination of intelligent parsing and supervised classification. We have further presented evidence that students who spend time crafting thoughtful responses show improved learning outcomes when practicing earlier material.

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 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, and many helpful discussions regarding this work.

6. REFERENCES

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

[2] 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.

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

[4] 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.

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

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

[7] J. Park. Constructive multiple-choice testing system. *British Journal of Educational Technology*, 41(6):1054–1064, 2010.