Predicting Reading Comprehension from Constructed Responses: Explanatory Retrievals as Stealth Assessment

Kathryn S. McCarthy[1(✉)], Laura K. Allen[2], and Scott R. Hinze[3]

[1] Georgia State University, Atlanta, USA
kmccarthyl2@gsu.edu
[2] University of New Hampshire, Durham, USA
laura.allen@unh.edu
[3] Middle Georgia State University, Macon, USA
scott.hinze@mga.edu

Abstract. Open-ended *constructed responses* promote deeper processing of course materials. Further, evaluation of these explanations can yield important information about students' cognition. This study examined how students' constructed responses, generated at different points during learning, relate to their later comprehension outcomes. College students (N = 75) produced self-explanations *during* reading and explanatory retrievals *after* reading. The Constructed Response Assessment Tool (CRAT) was used to analyze these responses across multiple dimensions of language and relate these textual features to comprehension performance. Results indicate that the linguistic features of post-reading explanatory retrievals were more predictive of comprehension outcomes than self-explanations. Further, these models relied on different indices to predict performance.

Keywords: Natural language processing · Science learning · Stealth assessment

1 Introduction

Learning from text is a critical skill, but many students struggle with content-based reading [1]. Prompting students to generate *constructed responses* (e.g., verbal protocols, summaries) is beneficial because it encourages active processing [2, 3] and these responses can also serve as "stealth assessments" [4, 5] of in situ learning that continually update a learner model and drive feedback without needing to wait for more formal checkpoint quizzes or module exams.

In the current study, we explore the use of *explanatory retrieval* prompts as stealth assessments. Explanatory retrievals are a type of constructed response in which students explain what they have just read from memory. As an elaborative or constructive version of retrieval practice, explanatory retrieval may yield superior comprehension as compared to free recall prompts or completing multiple-choice or fill-in-the-blank tests

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 197–202, 2020.
https://doi.org/10.1007/978-3-030-52240-7_36

[6, 7]. Not only is this approach effective, but it is also practical in the sense that asking students to "explain what you have just read about [*topic*]" rather than answer a series of quiz questions reduces the need for instructors or instructional designers to generate numerous items. Finally, these activities may have value as stealth assessments that can track students' learning processes and progress.

Although explanatory retrievals are beneficial for learning, they are often underutilized in the classroom due to the arduous nature of scoring open-ended responses [8]. Fortunately, natural language processing (NLP) tools have afforded an increased use of constructed responses within educational technologies [9, 10]. NLP analyses can be used to automate scoring and provide targeted feedback for a variety of constructed responses including think-alouds [11], self-explanations [12], summaries [13, 14], and essays [15]. Notably, the indices implicated in these analyses vary across constructed response type, presumably because they reflect different strategies and cognitive processes. Taken together, this research demonstrates the potential for analyzing explanatory retrievals as a mode of stealth assessment, but also highlights the need to consider how explanatory retrievals might differ from other forms of constructed response.

Thus, in the current study, we examine how linguistic features of explanatory retrievals (ERs) relate to comprehension test performance. We also examine how ERs compare with another type of constructed response, self-explanation (SE), for which linguistic features have been studied. The prior research guides two primary hypotheses: 1) The linguistic features of the responses will provide information predictive of subsequent comprehension test performance and 2) The features of ERs that predict comprehension performance will differ from the predictive features in SEs. In other words, as a retrieval (i.e., memory-based) process, post-reading ER may bring to bear different strategies and processes than what is found in concurrent SEs.

2 Method

2.1 Design and Procedure

College students (N = 75; M_{age} = 25.04; 72% female; 13% ESL) read two science texts. At nine points in each text, students were directed to generate an SE. After reading, participants were prompted to produce an ER. The instructions specified the goal was not to simply recall as much as possible, but to provide a coherent explanation of the information in the text. After reading and explaining both texts, participants completed multiple-choice comprehension tests for each text. Each test included four memory items and four inference items.

2.2 Data Processing

SEs were combined to create an "aggregated SE" for each text [16–18]. These aggregated SEs and the ERs were submitted to the Constructed Response Analysis Tool (CRAT) [19]. CRAT calculates more than 700 indices related to 1) similarities (key words overlap, latent semantic analysis) between a source text and a constructed

response and 2) lexical sophistication and text properties. After the SEs and ERs had been analyzed by CRAT, the dataset was reduced based on multicollinearity and relation to the dependent variable. Thus, when two variables were highly multicollinear ($r > .70$), only the index most strongly related to the dependent variable was retained. Additionally, indices that exhibited a weak or absent relationship with the dependent variable ($r < .10$) were removed from the dataset. After this process, there were 50 CRAT indices remaining for the machine learning analyses.

2.3 Supervised Classification and Validation

Supervised machine learning techniques were used to predict students' comprehension scores. *Caret* for R [20] was used to train Linear Regression, Support Vector Machine (SVM), and Random Forest models. All models were evaluated using leave-one-out cross-validation (LOOCV) in which $k - 1$ instances were used in the training set and the model was tested on the instance not used in the training data. This process was repeated k times until each instance was used as the test set. LOOCV develops models that are more generalizable when applied to new data.

3 Results

On average, students' aggregated SEs contained 172.69 (SD = 99.75) words, whereas their ERs contained 90.97 (SD = 45.52) words. Word count was included as a control variable in our models; however, it was not an important feature of any of the models.

The response types (SE, ER) were tested independently using the same regression algorithms (Linear Regression, SVM, Random Forest). A summary of model accuracies is presented in Table 1. Overall, the SVM performed the best for both SE and ER data. The CRAT indices accounted for 15% (SE) and 25% (ER) of variance in comprehension scores, suggesting that the properties of the retrievals were more informative of students' comprehension of text content.

Table 1. Description of model accuracy.

	Self-Explanation (SE)		Explanatory Retrieval (ER)	
Algorithm	RMSE	R^2	RMSE	R^2
Linear Regression	1.97	0.04	1.76	0.12
SVM (Polynomial)	**1.67**	**0.15**	**1.52**	**0.25**
Random Forest	1.67	0.13	1.50	0.24

To more closely examine the CRAT indices driving the model predictions, we examined the scaled variable importance of indices in the SVM models. Four of the top five variables in the SE model were adjective keywords from the COCA corpus. They related to *academic adjective keywords, magazine adjective keywords, fiction adjective keywords, news adjective keywords,* and *academic bigram keywords.* In comparison,

the top five variables in the ER model were *academic bigram keywords, word imageability, academic keywords, age of acquisition for content words, and fiction keywords.* These results indicate that the descriptive content (i.e., adjectives) of the SEs were most predictive of comprehension scores, whereas the ERs were related to a wider variety of textual information, particularly lexical sophistication.

4 Discussion

This study examined the potential of explanatory retrievals (ERs) to serve as a form of stealth assessment of reading comprehension performance. Given that open-ended retrieval attempts can vary widely in quality [7], automating the evaluation of ER practice can make it more feasible to include ER tasks in the classroom. This study demonstrated modest, but promising results. In particular, our best model (SVM Polynomial) accounted for 15% and 25% of the variance using the properties of SEs and ERs, respectively. These results support the extant work demonstrating that natural language processing techniques can be used to model important comprehension processes [11–15].

A more novel finding in this present study is that, as predicted, different types of constructed responses were not uniformly related to reading comprehension performance. That is, SE responses and ER responses relied on some different features to predict comprehension and did so to different degrees of success. This supports the idea that different constructed responses influence and predict comprehension in different ways. Further work will more closely examine these different linguistic features in context to understand *why* different types of linguistic features are more or less predictive in a particular type of response and how these different processes impact different aspects of learning (i.e., memory vs. inference and application). The goal of this study was to compare and contrast across types of constructed responses and how each might provide different insights into learning processes. However, in future work, we plan to leverage the unique contributions of both in a combined model in which features of SEs and ERs are used to predict performance.

One limitation of note is that LOOCV was conducted at the item level, with the same participants generating multiple items. Further research with larger data sets will examine how these models generalize to entirely independent datasets. In addition, this study relied only on the CRAT tool to analyze linguistic features of the constructed responses. Existing work on analysis of constructed responses [15–18] suggests that our models will have higher accuracy if they include indices that characterize text across multiple dimensions (e.g., lexical, syntax, cohesion). Thus, future work will examine the value of employing additional linguistic analysis tools to account for variance in other dimensions of language.

Overall, the results of this study suggest that ERs can serve as both powerful learning activities and as assessments of developing comprehension. However, more work is needed to improve and refine automated procedures for scoring and providing feedback based on these responses. The ultimate goal of this research is to use these linguistic indices to facilitate nuanced assessments of constructed responses that can drive improved formative feedback and personalization in educational technologies.

Acknowledgements. This research was made possible in part by grants from the Spencer Foundation (201900217) and the Institute for Education Sciences (R305A190063 and R305A180261). The views expressed are those of the authors and do not necessarily reflect the views of the funding agencies.

References

1. Goldman, S.R., Snow, C.E.: Adolescent literacy: development and instruction. In: Polatsek, A., Treiman, R. (eds.) Handbook on Reading, pp. 463–478. Oxford University Press, New York (2015)
2. Bertsch, S., Pesta, B.J., Wiscott, R., McDaniel, M.A.: The generation effect: a meta-analytic review. Mem. Cogn. **35**(2), 201–210 (2007)
3. McNamara, D.S.: Reading Comprehension Strategies: Theories, Interventions, and Technologies. Psychology Press, Hove (2007)
4. Shute, V.J.: Stealth assessment in computer-based games to support learning. Comput. Games Instr. **55**(2), 503–524 (2011)
5. Shute, V.J., Kim, Y.J.: Formative and stealth assessment. In: Spector, J.M., Merrill, M.D., Elen, J., Bishop, M.J. (eds.) Handbook of Research on Educational Communications and Technology, pp. 311–321. Springer, New York (2014). https://doi.org/10.1007/978-1-4614-3185-5_25
6. Endres, T., Carpenter, S., Martin, A., Renkl, A.: Enhancing learning by retrieval: enriching free recall with elaborative prompting. Learn. Instr. **49**, 13–20 (2017)
7. Hinze, S.R., Wiley, J., Pellegrino, J.W.: The importance of constructive comprehension processes in learning from tests. J. Mem. Lang. **69**(2), 151–164 (2013)
8. Hinze, S.R., Wiley, J.: Testing the limits of testing effects using completion tests. Memory **19**(3), 290–304 (2011)
9. Crossley, S.A., McNamara, D.S.: Adaptive Educational Technologies for Literacy Instruction. Routledge, New York (2016)
10. Passonneau, R.J., McNamara, D.S., Muresan, S., Perin, D.: Preface: special issue on multidisciplinary approaches to AI and education for reading and writing. Int. J. Artif. Intell. Educ. **27**(4), 665–670 (2017)
11. Magliano, J.P., Millis, K.K., Levinstein, I., Boonthum, C.: Assessing comprehension during reading with the Reading Strategy Assessment Tool (RSAT). Metacogn. Learn. **6**(2), 131–154 (2011)
12. Jackson, G.T., McNamara, D.S.: Applying NLP metrics to students' self-explanations. In: Applied Natural Language Processing: Identification, Investigation and Resolution, pp. 261–275. IGI Global (2012)
13. Kim, M.K., Gaul, C.J., Kim, S.M., Madathany, R.J.: Advance in detecting key concepts as an expert model: using student mental model analyzer for research and teaching (SMART). Technol. Knowl. Learn. 1–24 (2019). https://doi.org/10.1007/s10758-019-09418-5
14. Li, H., Cai, Z., Graesser, A.C.: Computerized summary scoring: crowdsourcing-based latent semantic analysis. Behav. Res. Methods **50**(5), 2144–2161 (2018)
15. Crossley, S.A., Roscoe, R., McNamara, D.S.: Predicting human scores of essay quality using computational indices of linguistic and textual features. In: Biswas, G., Bull, S., Kay, J., Mitrovic, A. (eds.) AIED 2011. LNCS (LNAI), vol. 6738, pp. 438–440. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21869-9_62

16. Varner, L.K., Jackson, G.T., Snow, E.L., McNamara, D.S.: Does size matter? Investigating user input at a larger bandwidth. In: Proceedings of the 26th International Florida Artificial Intelligence Research Society (FLAIRS) Conference, pp. 546–549. AAI Press (2013)
17. Allen, L.K., McNamara, D.S.: You are your words: modeling students' vocabulary knowledge with natural language processing. In: Proceedings of the 8th International Conference on Educational Data Mining (EDM), pp. 258–265. Madrid, EDM (2015)
18. Allen, L.K., Snow, E.L., McNamara, D.S.: Are you reading my mind? Modeling students' reading comprehension skills with natural language processing techniques. In: Baron, J., Lynch, G., Maziarz, N., Blikstein, P., Merceron, A., Siemens, G. (eds.) Proceedings of the 5th International Learning Analytics & Knowledge Conference (LAK 2015), pp. 246–254. ACM, Poughkeepsie (2015)
19. Crossley, S.A, Kyle, K., Davenport, J., McNamara, D.S.: Automatic assessment of constructed response data in a chemistry tutor. In: Proceedings of the 9th International Educational Data Mining (EDM) Society Conference, pp. 336–340. EDM (2016)
20. Kuhn, M., et al.: Package 'caret'. R J. (2020)