Guerrero, T. A., & Wiley, J. (2019). Using “Idealized Peers” for Automated Evaluation of Student Understanding in an Introductory Psychology Course. In *Proceedings of the 20th International Conference on Artificial Intelligence in Education* (pp. 133-143). Springer, Cham. https://doi.org/10.1007/978-3-030-23204-7_12

Using “idealized peers” for automated evaluation of student understanding in an introductory psychology course

Tricia A. Guerrero and Jennifer Wiley

University of Illinois at Chicago, Chicago IL 60647, USA
tguerr9@uic.edu, jwiley@uic.edu

Abstract. Teachers may wish to use open-ended learning activities and tests, but they are burdensome to assess compared to forced-choice instruments. At the same time, forced-choice assessments suffer from issues of guessing (when used as tests) and may not encourage valuable behaviors of construction and generation of understanding (when used as learning activities). Previous work demonstrates that automated scoring of constructed responses such as summaries and essays using latent semantic analysis (LSA) can successfully predict human scoring. The goal for this study was to test whether LSA can be used to generate predictive indices when students are learning from social science texts that describe theories and provide evidence for them. The corpus consisted of written responses generated while reading textbook excerpts about a psychological theory. Automated scoring indices based in response length, lexical diversity of the response, the LSA match of the response to the original text, and LSA match to an idealized peer were all predictive of human scoring. In addition, student understanding (as measured by a posttest) was predicted uniquely by the LSA match to an idealized peer.

Keywords: Automated assessment, Natural language processing, Latent semantic analysis, Write aloud methodology

1 Introduction

1.1 Generative Activities

Teachers may wish to use open-ended learning activities and tests, but they are burdensome to assess compared to forced-choice instruments. At the same time, forced-choice assessments suffer from issues of guessing (when used as tests) and may not encourage valuable behaviors of construction and generation of understanding (when used as learning activities). The use of generative learning activities such as prompting students to write explanations has been shown to be beneficial to improving understanding when learning in science [1,2,3,4]. Generating explanations can prompt students to engage in the construction of a mental model of the concepts in the text. The process of writing explanations may be effective because it prompts students to generate inferences and make connections across the text and to their own prior knowledge.

Prior work has shown that engaging in constructive learning activities, such as generating explanations, increases student understanding compared to other more passive activities such as re-reading [3]. However, other work suggests that the quality of the explanations that are generated may matter [2,5]. This means that students may need feedback on the quality of their explanations in order to gain the benefits of engaging in this learning activity. In turn, this then places a large burden on teachers. However, if evaluation of student responses such as explanations could be accomplished using automated natural language processing indices, then teachers could utilize open-ended learning activities with increased frequency. And, the same methods could also be used to score open-ended test questions.

1.2 Using latent semantic analysis in automated evaluation of responses

Latent semantic analysis (LSA) has been useful in automated evaluation of constructed student responses as it can be used to generate an index representing the overlap in semantic space between two texts [6]. Foltz et al. [7] used multiple approaches with LSA to assess short-answer essays written about a cognitive science topic: how a particular connectionist model accounts for a psycholinguistic phenomenon (the word superiority effect). Measures of semantic overlap were obtained by comparing student essays to the original text in two ways: one using the whole text and one using selected portions that were deemed most important. Both approaches were found to be highly correlated with scores obtained from human graders who coded for content and quality of writing. Similarly, Wolfe et al. [8] derived LSA scores by comparing short student essays about heart functioning to a standard textbook chapter, and found these LSA scores predicted the grades assigned by professional graders (using a 5-point holistic measure of quality) as well as the scores that students received on a short-answer test of their knowledge of the topic.

In addition to comparing student responses to the original text or a standard text, another approach has compared student responses to an expert summary. León et al. [9] had students read either a narrative excerpt from a novel (The Carob Tree Legend) or an encyclopedia entry (The Strangler Tree) and write a short summary. The LSA comparison to the "gold standard" expert response was more predictive of human scoring than the LSA comparison to the original text. Similar results have been obtained in studies with students writing about ancient civilizations, energy sources and the circulatory system [10], and in response to conceptual physics problems [11].

Prior research has used LSA to make comparisons between student responses and expert responses; however, when experts write responses they tend to use more academic language and make different connections and elaborations than students based on their prior knowledge [12]. Thus, researchers have also explored making comparisons to peer responses. Both Foltz et al. [7] and León et al. [9] used exact responses written by peers to compute an average LSA score from comparisons of each student response with all other student responses. These average scores were predictive of human scoring. Other studies have used LSA to contrast student responses against "best peer" responses. Ventura et al. [12] had students write responses to conceptual physics problems within an intelligent tutoring system. Student responses were compared to both an expert response and a best peer response. The best peer response was taken randomly from all responses given the grade of an A. When comparing the LSA match to the expert response and the best peer response, the LSA match to the best peer more accurately predicted the letter grade assigned by a human grader.

Other work has used LSA measures based in "idealized" peer comparisons to predict not just human coding, but also student understanding. In Wiley et al. [13], students read texts as part of a multiple document unit on global warming, and were asked to generate an explanation about how global warming occurs. An idealized peer response was constructed to include the key features from the best student essays. The LSA scores obtained by comparing the student responses to the idealized peer response were predictive of both holistic human scoring, as well as student understanding as measured by an inference verification test given at the end of the unit.

The main goal for the present research was to further explore the effectiveness of automated scoring using peer-based LSA measures to predict understanding from a social science text in which a theory was presented along with supporting empirical research and examples to explain the theory. This text structure is representative of the style of many social science textbooks, including those in introductory psychology. With such texts, it is the responsibility of the reader to understand how and why the cited studies and examples support the theory as described. The present study tested whether the LSA match between student comments generated while reading and an experimenter-constructed idealized peer could serve not only as a predictor of holistic human coding, but also serve as a measure of student understanding.

2 Corpus and Human Scoring of Responses

2.1 Corpus

The corpus consisted of short written responses generated by 297 undergraduates while reading a text about cognitive dissonance, a key topic that is generally covered in most courses in introductory psychology. The comments were written by undergraduate students in an introductory psychology course (188 females; Age: *M* = 18.93, *SD* = 1.16) as a part of a homework assignment administered through the Qualtrics survey platform. All responses were edited to correct any typographical errors as well as to expand contractions and abbreviations. The textbook excerpt that was assigned for this topic had a Flesch-Kincaid reading level of 12.5 and contained 863 words in 5 paragraphs. The excerpt began with a real-world example followed by a description of the theoretical concept. The passage then described two research studies which provided empirical support for cognitive dissonance theory. Students were given an initial opportunity to read this textbook excerpt in an earlier homework assignment. During the target activity for this study, students were given a brief instructional lesson on how to generate explanations to support their learning from text:

> *As you read the texts again today, you should try to explain to yourself the meaning and relevance of each sentence and paragraph to the overall purpose of the text. At the end of each sentence and paragraph, ask yourself questions like:*
>
> - *What does this mean?*
> - *What new information does this add?*
> - *How does this information relate to the title?*
> - *How does this information relate to previous sentences or paragraphs?*
> - *Does this information provide important insights into the major theme of the text?*
> - *Does this sentence or paragraph raise new questions in your mind?*

Students then saw an example text with associated example responses to these questions that could be written at various points in the text.

After the lesson, students reread the textbook excerpt on cognitive dissonance. At the end of each of the 5 paragraphs, they were prompted to "*write your thoughts*" for the current section of the text similar to a "type-aloud" or "write-aloud" procedure [14]. In addition, they were asked to write their thoughts at the end of the entire text. They were reminded to think about the questions given in the instructions which were present in a bulleted list on the screen as a reference while they wrote their thoughts. The 6 thought statements were concatenated into a single response for each student with an average length of 190 words (*SD* = 114, range: 6 - 728) and an average lexical diversity of 58.05 (*SD* = 34.71, range: .01 - 125.50).

Several additional measures were available for each student. Student understanding of the topic following the homework activity was measured by performance on a 5-question multiple-choice comprehension test (*M* = 2.44, *SD* = 1.21). As seen in Table 1, these questions were designed to test the ability to reason from information in the text, and to construct inferences about information left implicit in the text, not just verbatim memory for facts and details. Students did not have access to the text while completing the test. This was collected during the next week's homework activity which served as a practice test for the upcoming exam. The data set also included measures of reading ability (ACT scores, *M* = 23.72, *SD* = 3.62) and prior knowledge (performance on a 5-item multiple choice pretest on the topic given during the first week of the course, *M* = 1.87, *SD* = 1.14). Prior studies [except 13] have generally not included reading ability as a predictor when using automated evaluation systems. This leaves open the question of whether automated evaluation systems are solely useful in predicting general reading ability (and detecting features of essays written by better readers) rather than predicting the quality of features in specific responses.

2.2 Human Scoring of Responses

Student responses were scored by two human coders using a rubric adapted from McNamara et al. [15] and Hinze et al. [2], similar to what a teacher might use to quickly assess their quality. A score of 0 was assigned to responses that represented little to no effort: consisting of only non-word gibberish ("dfkashj"), two or fewer words per paragraph, or only verbatim phrases that were copied and pasted from the original text. Responses that included paraphrased ideas from the text (but no additional elaborations) were assigned a 1 (e.g., “Possible ways to reduce cognitive dissonance include changing one’s behavior,” “Two scientist managed an experiment cognitive dissonance with children and their toys”). Responses that showed evidence of constructive processing, such as when students identified connections not explicit in the text, were assigned a score of 2. This could occur through identifying the relations between theories and evidence, or making connections to relevant prior knowledge (e.g., “Whenever people have conflicting beliefs and actions, some sort of resolution must occur. The conflict causes psychological distress and must be removed. In order to reduce cognitive dissonance, they must alter their beliefs to match the action or altering behaviors to match the belief”). Interrater agreement between two coders resulted in Cohen’s kappa of .92.

Table 1. Paragraph 3 of cognitive dissonance text, idealized-peer response from concepts appearing in highest scoring student responses, and example test question.

Text Excerpt	
In 1959, Festinger and Carlsmith conducted an experiment which tested cognitive dissonance theory. Participants were asked to spend an hour performing a very boring task…. These participants were asked to recommend the experiment they had just completed to other potential participants who were waiting to complete the experiment. They were instructed to tell these potential participants that the experiment was fun and enjoyable. Half of the participants in this group were paid $1 to recommend the experiment and the other half were paid $20. These participants were then taken to the interview room and asked the same questions as the participants in the control group, who were not paid and were not asked to talk to other participants. The participants in the $20 group responded similarly to the participants in the control group, namely that they did not find the experiment to be enjoyable and that they would not sign up to participate in a similar experiment. In contrast, participants in the $1 group rated the experiment as more enjoyable than participants in the other two groups, and indicated that they would be more willing to participate in another similar experiment.	
Most frequent concepts in best responses	Idealized-peer response
-Identify groups performing similarly (18%) -Question the reasoning for results of study (72%)	The control group and the $20 group both told the truth that they did not enjoy the experiment. The $1 group rated the experiment as more enjoyable. This does not make sense. Why would the $1 group say it was fun?
Test Question	
Imagine that the theory in the text was incorrect and that people do not experience cognitive dissonance. Which result of the Festinger experiment (about getting paid to do a boring task) would you expect? a. The control group who got paid nothing would have said they found the task very interesting. b. The group paid $1 would have said they found the task to be boring. c. The group paid $20 would have said they found the task to be very interesting. d. How much people got paid would not have had a bigger effect on what they said about the task.	

2.3 Idealized-peer response

The idealized-peer response was constructed by selecting concepts and phrases that appeared most frequently in responses to each of the 5 paragraphs across the best student comments (i.e., scored as “2” by human raters). An example of the idealized peer response for one paragraph is shown in Table 1. The idealized response, written at the 8th grade level, included a paraphrase of the main point and 1-2 of the most frequent elaborations for each paragraph. The elaborations were often written in the first and second person. Elaborations also included explicit connections between the theories presented and the experiments that were left implicit in the original text, and metacognitive comments (e.g., I am not sure why they would do that?).

3 Results using Automated Scoring Indices

3.1 Automated scoring indices

Four automated measures were computed. Two measures were calculated using LSA. The first compared the student response to the actual text excerpt that was read (LSAORIG). The second compared the student response to the idealized-peer response (LSAIDEAL). In addition, the total response comment length (LENGTH) was computed using Linguistic Inquiry and Word Count (LIWC) [16] and the lexical diversity (LEXDIV) of all words in each student response was measured using Coh-Metrix index LDVOCDa [17]. The length of a response is often predictive of human scoring, accounting for over 35% of the variance in human-scored responses [18-20]. The variety of words used can also predict human scoring. In essays where students were asked to describe the popularity of comic books or wearing name-brand fashions, or to write letters responding to a complaint or welcoming an exchange student, the lexical diversity of the response was a positive predictor of essay grades assigned by human raters [20]. While features such as the number and diversity of words within a student response may influence human scoring, other work has found that length may not predict student understanding, and the relation between lexical diversity and understanding may became negative once the LSA match with the idealized peer essay is taken into account [13]. To further explore these relations, two additional automated measures (LENGTH, LEXDIV) were included in the present analyses.

Table 2. Correlations among measures for student responses.

	HUMAN	LENGTH	LEXDIV	LSAORIG	LSAIDEAL
LENGTH	.46**	-			
LEXDIV	.55**	.54**	-		
LSAORIG	.68**	.55**	.55**	-	
LSAIDEAL	.79**	.50**	.54**	.83**	-
POSTTEST	.15**	.10	.08	.19**	.23**

**Correlations are significant at the 0.01 level.

As shown in Table 2, human scoring (HUMAN) predicted posttest performance (POSTTEST). LSA measures predicted human scoring, and were at least as strong of predictors of posttest performance as human scoring. Descriptively, the strongest single predictor of posttest performance was the match with LSAIDEAL (although this correlation was not significantly stronger than the correlation with HUMAN scoring, $z = 1.01$, $p = .16$). Despite the significant correlations among measures, variance inflation factors in all reported analyses remained below 1.8 indicating that multicollinearity was not an issue for analyzing the measures together in regressions.

3.2 Relation of automated scoring to human scoring

As shown in Table 2, the simple correlations between human scores and all four automated measures were significant. However, as shown in Table 3, when they were all entered simultaneously into a regression model, LSAORIG was no longer a significant predictor of human scoring. LSAIDEAL and LEXDIV both remained as positive unique predictors of the human scores, with the full model accounting for 58% of the variance in human scores, $F(4,292) = 130.53$, $p < .001$.

Table 3. Human-scored quality as predicted by automated measures.

Variable	Unstandardized Beta (*B*)	Std. Error	Standardized Beta (*β*)	*t*-value	*p*-value
(Constant)	.27	0.08		3.17	.002
LENGTH	0.00	0.00	.04	0.814	.42
LEXDIV	0.00	0.00	.16	3.56	< .001
LSAORIG	-0.03	0.22	-.01	-0.15	.88
LSAIDEAL	2.68	0.25	.69	10.72	< .001

3.3 Relation of automated scoring to student understanding

As shown in Table 2, the simple correlations between student understanding (assessed by posttest scores) and automated measures were only significant for the two LSA measures (LSAORIG and LSAIDEAL). Posttest scores were not significantly predicted by response length (LENGTH) or lexical diversity (LEXDIV). Further, as shown in Table 4, only LSAIDEAL remained as a significant predictor, $R^2 = .04$ $F(4,292) = 4.13$, $p = .003$, when all 4 automated measures were entered simultaneously.

Table 4. Student understanding as predicted by automated measures.

Variable	Unstandardized Beta (*B*)	Std. Error	Standardized Beta (*β*)	*t*-value	*p*-value
(Constant)	1.23	0.31		4.01	< .001
LENGTH	0.00	0.00	.00	0.02	.99
LEXDIV	0.00	0.00	-.06	-0.79	.43
LSAORIG	0.13	0.82	.02	0.16	.88
LSAIDEAL	2.16	0.93	.24	2.33	.02

3.4 Unique contribution of LSAIDEAL over and above reader characteristics

It is typically the case that students who are better readers or who have prior knowledge of a topic will develop better understanding when learning from text. Indeed, both ACT scores (r = .25) and prior knowledge measures (PRETEST, r =. 29) were significant predictors of posttest scores. However, as shown in Table 5, LSAIDEAL remained as a significant predictor even when both ACT scores and prior knowledge were included in the model, R^2 = .17, F(3,249) = 16.79, p < .001.

Table 5. Student understanding as predicted by LSAIDEAL and reader characteristics.

Variable	Unstandardized Beta (*B*)	Std. Error	Standardized Beta (*β*)	*t*-value	*p*-value
(Constant)	-0.80	0.54		-1.49	.14
ACT	0.07	0.02	.22	3.77	< .001
PRETEST	0.22	0.06	.21	3.63	< .001
LSAIDEAL	1.97	0.49	.23	4.01	< .001

3.5 Comparison of LSAIDEAL to other LSA alternatives

There are several possible reasons why idealized peer responses were more predictive of understanding than the original text. One may be that sections in introductory textbooks contain a large number of ideas about each topic. The idealized peer response may gain its power by selecting out the most relevant ideas from the section. Thus, when a student's response overlaps heavily with the content of the idealized peer response, this may reflect that student's ability to identify, select, and attend to the most relevant features of the text. This may be similar to the predictive value of just the most important sentences within the text [7]. A second possible reason may be because idealized peer comments are written in more colloquial language that other students may be more likely to use [12,13]. A third possible reason is that idealized peer responses may explicitly mention key inferences and connections that are left implicit in the text [12]. And finally, constructing an idealized peer response from multiple high-quality student responses may be better than using only one randomly selected "best student" because comments vary and contain many idiosyncrasies that may be relevant based on the prior knowledge of one individual more so than another.

To better understand what may be responsible for the predictive power of the idealized peer response, several alternative LSA comparisons were computed: the match of each student's comments to the same concepts in the LSAIDEAL but written in academic language at a 12th grade level (ACADEMIC), to an automated selection (selected by R package LSAfun [22]) of the important sentences in each section of the text (LSAFUN), to important sentences as selected by expert (SELECTED), to sentences written by an expert to represent the explicit connections that need to be made to comprehend the text (EXPLICIT), and to a randomly chosen single best peer response (BESTPEER). The partial correlations after controlling for the unique contributions to prediction from reading ability and prior knowledge are shown in Table 6.

Table 6. Partial correlations among LSA measures and student understanding.

	Posttest
LSAORIG	.20**
LSAIDEAL	.25**
ACADEMIC	.24**
LSAFUN	.22**
SELECTED	.25**
BESTPEER	.23**
EXPLICIT	.23**

**Partial correlations are significant at the 0.01 level.

Note. Controlling for reading ability (ACT) and prior knowledge (pretest).

4 Discussion

This study tested multiple automated measures that may be useful for assessing student understanding. Students wrote responses while reading a textbook excerpt on cognitive dissonance, a commonly taught subject in introductory psychology courses. All responses were scored for quality by both humans and using automated measures.

Although lexical diversity of the comments was a significant positive predictor of human scoring, it was not predictive of student understanding as measured by the posttest. When the intended purpose of a learning activity is to promote student understanding, and when the goal for using automated measures is to predict student understanding (rather than to match holistic impressions of human scorers), then features such as length and lexical diversity may be less useful.

In contrast, the LSA match with the idealized-peer response provided a better fit for both human scoring and for student understanding than did the LSA match to the original text. Although this predictive model accounted for a relatively small proportion of the variance in test scores, it provides a first step in exploring how learning activities that prompt students to record their thoughts online as they are attempting to comprehend a text might be able to utilize automated evaluation techniques.

This study represents an advance beyond prior work by the inclusion of reading ability and prior knowledge in the prediction models, as well as by testing across a wide range of LSA metrics. Similar results were seen between idealized responses written in academic and more colloquial language indicating that the use of peer language may not be as important as hypothesized. Further, the use of idealized peer responses that included multiple elements from several of the best students seemed to produce a better standard than a single randomly chosen best response (although this finding may be highly variable based on the single response chosen). Additionally, an expert may choose slightly better sentences than an automated system (LSAfun), but the advantage of automation may be important for broader implementation.

Another limitation of the current implementation was that the student responses needed to be edited to correct misspellings and abbreviations prior to processing to achieve these results. However, simply requiring students to use a spelling and grammar check tool prior to submission has been successful in properly editing responses for processing [10]. Adding that feature could also aid automation in this case.

5 Conclusion and Future Directions

The main goal for the present research was to further explore the effectiveness of automated scoring using LSA to predict understanding from a social science text in which a theory was presented along with supporting empirical research and examples to explain the theory. The results of the present study demonstrated that the LSA match between student comments and an idealized peer could serve not only as a predictor of holistic human coding, but also as a measure of student understanding.

Ultimately, the motivation behind developing and testing for effective means of automated coding of student responses is to enable the development of automated evaluation and feedback systems that support better student comprehension when attempting to learn from complex social science texts. Generative activities can be beneficial for learning, but they may be especially effective when feedback is provided to students. Moving forward, the next step in this research program is exploring how this automated scoring approach can be used to provide intelligent feedback to students as they engage in these learning activities.

Though the predictive power of this approach is limited, the results of the present study are promising as they suggest that evaluations of response quality derived from an LSA index based in the match between students' comments and an idealized-peer might be just as helpful as having a teacher quickly assess the quality of student comments made during reading. Utilizing these automated measures may make it more feasible for teachers to assign learning activities that contain open-ended responses, and for students to learn effectively from them.

Acknowledgements. This research was supported by grants from the Institute for Education Sciences (R305A160008) and the National Science Foundations (GRFP to first author). The authors thank Grace Li for her support in scoring the student responses, and Thomas D. Griffin and Marta K. Mielicki for their contributions as part of the larger project from which these data were derived.

References

1. Chi, M. T. H.: Self-explaining expository texts: The dual process of generating inferences and repairing mental models. In: Glaser, R (ed.) Advances in instructional psychology, pp. 161-237. Erlbaum: Mahwah, NJ (2000).
2. Hinze, S. R., Wiley, J., Pellegrino, J. W.: The importance of constructive comprehension processes in learning from tests. Journal of Memory and Language, 69, 151-164. (2013).
3. Chi, M. T. H., de Leeuw, N., Chiu, M. H., LaVancher, C.: Eliciting self-explanation improves understanding. Cognitive Science, 18, 439-477 (1994).
4. McNamara, D. S.: SERT: Self-explanation reading training. Discourse Processes, 38, 1-30 (2004).
5. Guerrero, T. A., Wiley, J.: Effects of text availability and reasoning processes on test performance. In: Proceedings of the 40th Annual Conference of the Cognitive Science Society, pp. 1745-1750. Cognitive Science Society, Madison, WI (2018).
6. Landauer, T. K., Foltz, P. W., Laham, D.: An introduction to latent semantic analysis. Discourse processes, 25, 259-284 (1998).
7. Foltz, P. W., Gilliam, S., Kendall, S.: Supporting content-based feedback in on-line writing evaluation with LSA. Interactive Learning Environments, 8, 111-127 (2000).
8. Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., Landauer, T. K.: Learning from text: Matching readers and text by latent semantic analysis. Discourse Processes, 25, 309-336 (1998).
9. León, J. A., Olmos, R., Escudero, I., Cañas, J. J., Salmerón, L.: Assessing short summaries with human judgments procedure and latent semantic analysis in narrative and expository texts. Behavior Research Methods, 38, 616-627 (2006).
10. Kintsch, E., Steinhart, D., Stahl, G., LSA Research Group: Developing summarization skills through the use of LSA-based feedback. Interactive learning environments, 8, 87-109 (2000).
11. Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z., Hu, X.: Using LSA in AutoTutor: Learning through mized initiative dialogue in natural language. Handbook of latent semantic analysis, 243-262 (2007).
12. Ventura, M. J., Franchescetti, D. R., Pennumatsa, P., Graesser, A. C., Hu, G. J., Cai, Z.: Combining computational models of short essay grading for conceptual physics problems. In: Lester,J. C., Vicari, R. M., Paraguacu, F. (eds.) Proceedings of the Intelligent Tutoring Systems Conference, pp. 423-431. Springer, Berlin (2004).
13. Wiley, J., Hastings, P., Blaum, D., Jaeger, A. J., Hughes, S., Wallace, P., Griffin, T. D., Britt, M. A.: Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science. International Journal of Artificial Intelligence in Education, 27, 758-790 (2017).
14. Muñoz, B., Magliano, J. P., Sheridan, R., McNamara, D. S.: Typing versus thinking aloud when reading: Implications for computer-based assessment and training tools. Behavior research methods, 38, 211-217 (2006).
15. McNamara, D. S., Boonthum, C., Levinstein, I. B., Millis, K: Evaluating self-explanations in iSTART: comparing word-based and LSA algorithms. Handbook of latent semantic analysis, 227-241 (2007).
16. Pennebaker, J. W., Booth, R. J., Boyd, R. L., Francis, M. E.: Linguistic inquiry and word count: LIWC 2015. LIWC.net, Austin, TX (2015).
17. Graesser, A.C., McNamara, D. S., Louwerse, M. M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. Behavior research methods, instruments, and computers, 36, 193-202 (2004).

18. Dikli, S.: An overview of automated scoring of essays. Journal of Technology, Learning, and Assessment, 5, 1-35 (2006).
19. Kobrin, J. L., Deng, H., Shaw, E. J.: The association between SAT prompt characteristics, response features, and essay scores. Assessing Writing, 16, 154-169 (2011).
20. Ferris, D. R.: Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. TESOL Quarterly, 28, 414-420 (1994).
21. Crossley, S. A., McNamara, D. S.: Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. Journal of Research in Reading, 35, 115-135 (2012).
22. Guenther, F., Dudschig, C., Kaup, B.: LSAfun: An R package for computations based on Latent Semantic Analysis. Behavior Research Methods, 47, 930-944 (2015).