UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Automated scoring of originality using semantic representations

Permalink
https://escholarship.org/uc/item/17t264nk

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 36(36)

ISSN
1069-7977

Authors
Harbinson, J. Isaiah
Haarman, Henk

Publication Date
2014

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Automated scoring of originality using semantic representations

J. Isaiah Harbison (iharbison@casl.umd.edu)
Center for Advanced Study of Language and Department of Psychology, University of Maryland
7005 52nd Avenue, College Park, MD 27642 USA

Henk Haarmann (hhaarmann@casl.umd.edu)
Center for Advanced Study of Language, University of Maryland
7005 52nd Avenue, College Park, MD 27642 USA

Abstract

Originality, a key aspect of creativity, is difficult to measure. We tested the relationship between originality and similarity in two semantic spaces: latent semantic analysis (LSA) and pointwise mutual information (PMI). Similarity in both spaces was negatively correlated with human judgments of originality of responses on a test of divergent thinking. PMI was correlated more strongly both with human judgments of similarity and human judgments of originality. In particular, the average PMI between two phrases was found to be the strongest predictor of phrase similarity and originality, even performing better than participants' self assessments of their originality.

Keywords: creativity; originality; semantic spaces; PMI; LSA

Scoring Originality

Current methods of scoring creativity assessments have drawbacks. Trained human raters, the gold standard for scoring creativity, require time for training, time to perform the scoring, and time for adjudication between raters. In addition, consistency between raters is often difficult to obtain due to the inherent subjectivity of the task. Automated scoring provides an alternative to human raters and can provide immediate scores, potentially useful for giving participants feedback and making it easier to incorporate creativity assessments in experiments. However, these methods are currently blind to the content of the response. Two such methods include Elaboration, a count of the total number of words in a response, and Fluency, the total number of responses. A participant could score highly on these measures by making the same long response many times.

The present research tests the potential of semantically informed automated scoring methods, automated methods that are not blind to content but instead capture some aspect of human judgment concerning similarity between responses. The goal is to develop a scoring method that captures human intuition but has the ability to provide instantaneous scoring for potential use as feedback to participants. To this end, we tested the ability of two computational methods of representing semantic content, latent semantic analysis (LSA) and pointwise-mutual information (PMI), to score the originality (or rarity) of responses on a creativity task. Both LSA and PMI have proved successful in predicting human judgments of similarity and there is some evidence that similarity could be used a measure of overall creativity (Forster & Dunbar, 2009). Here we continue our examination of the potential of LSA and PMI to predict individual components of creativity. Previously, we found that the distance between a participant's responses could be used to predict that participant's flexibility, the number of responses categories included in their output (Blok, Harbison, Haarmann, Bloodgood, & Berens, 2011). The current plan is to use the distance between the responses of all participants relative to a common point of comparison as a measure of originality.

We will first discuss LSA and PMI and test their ability to account for human judgments of individual word similarity and phrase similarity. LSA has a standard method for representing phrases, but PMI does not. Therefore, we tested three different methods of phrase representation with PMI. The final step was testing the spaces against the originality data using similarity as the measure of originality.

Semantic Representations and Originality

For over a decade, methods of using the context of word use to create semantic representations have proved to be predictive of a variety of human intuitions concerning word meaning. Latent Semantic Analysis (LSA), the predominant method of generating these representations, has been successfully applied to a wide variety of material: from predicting human judgments of word similarity, to performing tests of English as a second language, to scoring essays (Landauer, Laham, & Foltz, 2003).

Particularly relevant for the present context, LSA has also been applied to evaluating creativity. Wang, Chang, and Li (2008) used LSA to grade responses to a creative problem solving task. Similar to uses of LSA for grading essays, they compared each response to several ideal responses to determine if the essays contained the relevant concepts. Forster and Dunbar (2009) used LSA to predict human judgments of creativity on the Uses of Objects Task and found that LSA similarity correlated at .60 with human judgments. Blok et al. (2011) examined LSA's ability to score a component of creativity, flexibility. Flexibility is how varied a participant's responses are from each other. We tested multiple methods of using the similarity of the participants' responses as a measure of flexibility and found that the distance in the semantic space did well predicting this aspect of creativity. The present work builds on that of Blok et al. (2011) by applying LSA to another key component of creativity: originality.

In contrast to LSA, PMI has not yet been applied to creativity data. However, it has shown promise, correlating more strongly with human data than LSA due to its ability to make use of larger copora. Both LSA and PMI create summary

statistics based on word use, creating matrices that track how often words occur within a specific context. With LSA, the number of times each word occurs within each document of corpus is tracked. Within the resulting matrix, the number of words represented in the semantic space determines the number of rows and the number of documents processed determines the number of columns. A new column is added with each new document. Though the number of columns are reduced through singular value decomposition (SVD) after the entire corpus has been processed, LSA places a relatively large demand on computational resources in order to create its word representations. In contrast, PMI uses a word-by-word matrix and the only words that need to be included are the target words (i.e., the words that will be included in a comparison). If there are 600 relevant words, then the PMI space is a 600 by 600 matrix no matter the number of documents that are processed. That is, the size of the word representations, even during initial processing, are not a function of the number of documents in the corpus. As a result, PMI can process much larger corpora than LSA and due to the use of larger corpora, it is able to outperform LSA (Bidiu, REF; Recchia & Jones, 2009) despite being a much simpler algorithm. Here we test if PMI can outperform LSA when predicting originality.

For both LSA and PMI, the choice of corpus and how a document is defined within the corpus is very important. For the purpose of this comparison, we used the Touchstone Applied Science Associates (TASA) corpus for LSA. The TASA corpus was created to represent the word exposure expected of individuals between kindergarten and the first year of college (Zeno, Ivens, Millard & Duvvuri, 1995) and has often been successfully used by LSA. We used the implementation of LSA found at lsa.colorado.edu. For PMI we made use of its ability to utilize larger corpora, using all articles of Wikipedia as of January 2011 as the corpus. We restricted the number of words in the space to non-stop words (e.g., short function words) present in the datasets below. This was a total of 10,461 words which created a 10,461 by 10,461 matrix. We defined "document" as a paragraph within an article. If a word appeared in the same paragraph as another word, their co-frequency would be incremented but not otherwise. This definition was slightly different from that of Recchia and Jones (2009), who also used Wikipedia as their corpus. They defined blocks of 10 consecutive sentences as their documents.

Word Representations

Words are represented within LSA as vectors, a row where the number of columns, originally the number of documents, is reduced through SVD to a number of columns determined by the individual creating the space. To determine the similarity of two words, the cosine between the two word vectors is calculated.

For PMI, words are also represented as vectors, where the number of columns is the total number of target words. However, to evaluate the similarity of two words, the vectors are not compared to each other, but instead the PMI between the two words is used. PMI is a measure of whether the joint probability of the co-occurrence of two words differs from what is expected by chance, with higher values of PMI indicating greater than chance co-occurrence. Specifically, the PMI between words i and j is:

$$pmi_{i,j} = log_2 \frac{p(i,j)}{p(i)p(j)}. \quad (1)$$

In the present work, we followed the example of Recchia and Jones (2009) and used a modified version of the standard PMI equation. Here the frequency of each word was used in place of its probability and the frequency of co-occurrence was used for the joint probability. We also removed the log transformation, so the resulting equation was:

$$pmi^*_{i,j} = \frac{freq(i,j)}{freq(i)freq(j)}. \quad (2)$$

Similarity Judgments and Semantic Space Predictions

Before testing the semantic spaces against the originality data, we tested their ability to predict human judgments of similarity for individual words and two-word phrases. This was done to assure that the spaces, particularly the PMI space—as the present LSA implementation has been thoroughly tested—properly reflect the underlying similarity between words, making the subsequent test of the ability of dissimilarity to predict originality a valid one. The phrase similarity test is particularly important to PMI as it has not often been tested for its ability to represent phrases.

Word Similarity

As an initial comparison, we tested both LSA and PMI on their ability to predict the human judgments of word similarity from the WAS353 dataset (Finkelstein et al., 2002). This dataset consists of judgments of similarity of 353 different word pairs, such as "coast"-"hill" and "rooster"-"voyage". Table 1 displays the Spearman rank correlations between the two semantic spaces and the human data. The new PMI space, at a correlation of 0.72, performed equivalently to the space created by Recchia and Jones (0.73; 2009) and better than LSA at 0.60. This suggests that the difference in the specification of the document, from 10 sentences to single paragraphs, did not undermine PMIs ability to predict word similarity.

Table 1: Correlation between semantic space word similarity and human judgment of word similarity.

	New PMI	Previous PMI	LSA
WS353	.72	.73	.60

Phrase Representation and Phrase Similarity

Phrases would appear to be a challenge for both semantic spaces as both spaces are created treating documents as bags of words, ignoring word order and even whether or not the words appear in the same sentence. Furthermore, both spaces consist of representations of individual words, not larger units of language, making it is necessary to combine the individual word representations to generate phrase representations. Despite this, LSA has been successful modeling larger language units with the centroid method described below. PMI has not been applied to phrase data as much, with the one exception being in predicting information search behavior (Fu & Pirolli, 2007). In the present study, we compared four such methods of phrase representation.

1. LSA Centroid

The centroid is the standard method for representing phrases with LSA. Each phrase is represented as a vector. For example, two words in a space with n columns would be represented as $W_1 = a_1, a_2, ..., a_n$ and $W_2 = b_1, b_2, ..., b_n$. The centroid of a phrase consisting of W_1 and W_2 would be represented as $a_1 + b_1, a_2 + b_2, ..., a_n + b_n$. The same method is used to represent paragraphs and even entire documents. To judge similarity, the cosine between the two phrases are taken, the same calculation that was used for comparing words.

2. PMI First Word (w1PMI)

As there are no current standards for how to represent phrases with PMI, we tested three different methods. The first served as a lower baseline for PMI phrase representation. It included only the first (non-stop) word from each phrase (w1PMI). Comparison against this single word representation method provides an indication of how much is gained by attempting to represent an entire phrase within PMI using the methods presented below. For example, if w1PMI does as well as the other methods, then nothing was gained by using the phrase representations.

3. Average Pairwise PMIs (avePMI)

As mentioned above, PMI representations allow for the comparison of individual words based on their relative probability of joint occurrence. One method of representing the relationship between two phrases is to take the average of these relationships between the words of the two phrases (avePMI). That is, we summed over all pairwise comparisons between the words of the two phrases and divided by the number of comparisons. This measure reflects how related the words of the two phrases are on average. This is similar to how PMI is used to calculate information scent, which is used to predict information search behavior (Fu & Pirolli, 2007).

4. Average Product PMI (prodPMI)

The last method tested here used a similar method to avePMI but instead of using the sum of all the pairwise word PMIs it used their product. This measure differed from avePMI in its sensitivity to smaller PMI values. The use of the product of the individual relationships has been successful when modeling retrieval when there is a contribution from both episodic memory and semantic memory (Kimball, Smith, & Kahana, 2007).

We tested these four methods of calculating phrase similarity against data collected by Mitchell and Lapata (2010). The rationale was the same for testing the new semantic space against the WS353 data: before using method of representation to test to something novel, such as its ability to capture originality judgments, it should be tested on some known data to establish that the spaces well represent the underlying phrase similarity. Neither LSA nor PMI have been previously tested against this data.

The phrase data consisted of the human rated similarity of 324 phrase pairs provided by 204 participants. Each phrase consisted of two words arranged into three phrase types: adjective-noun (e.g., public building-central authority), compound nouns (e.g., study group-computer company), and verb-object (e.g., pass time-cross line). There were high, medium, and low similarity phrase pairs with 36 at each level. Table 2 shows the results from the four methods tested as well as the best previous correlations with this data set (Blacoe & Lapata, 2012).

Table 2: Correlation between human judgment and automated methods of evaluating phrase similarity.

Phrase type	Adj-noun	Compound noun	Verb-object
LSA Centroid	.542	.648	.441
w1PMI	.516	.546	.544
avePMI	**.661**	**.695**	**.545**
prodPMI	.291	.670	.374
Best Prior	.48	.50	.35

The similarity between phrases was significantly correlated with the human judgments of similarity for each method of representation. The avePMI representations had the strongest correlation with human data for all three types of phrases, ranging from .545 to .695. Perhaps most surprising result was how well the w1PMI representations performed. Simply using the first word of each phrase to represent each phrase was sufficient to outperform the best performing method tested by Balcoe and Lapata (2012) and to tie the best method tested in the present study for the Verb-Object phrase type. The LSA Centroid method, while performing worse than avePMI, consistently performed better that the best prior results as well.

Given the single word and phrase similarity results, we felt confident that the different methods capture some aspect of human similarity judgments and can be used to test the ability of similarity to predicting judgments of originality.

Scoring Originality

The originality data was taken from an experiment testing the potential of brain wave entrainment to improve divergent thinking (Haarmann et al., 2011). The task consisted of a set of scenarios and instructions for participants to generate as many responses as possible and as creative responses as possible. The scenarios examined here were:

1. A light in the darkness
2. Cloth in the breeze

Note that the experiment included two other scenarios that were not examined: "Person A is lying down, person B is sitting and person C is standing" and "Person A walks, Person B jumps". These were excluded because it was thought they were too inherently compositional. After performing the task, participants were asked to rate their own responses, producing a self-rating. In addition, six external raters, individuals not taking part in the experiment, rated the originality of each response. For the analyses below, we used the average of the external raters to determine the External Rater score for an individual response.

Similarity and Originality

To apply the methods of representing phrases to the originality data, it was necessary to determine the point of comparison. That is, to what should each response be compared to determine its originality? With the phrase data, we directly compared the phrases to each other to determine their similarity. The most straight forward method, would be to compare each response to the scenario they are responses to. For example, given the response "a lightning flash in the sky", the non-stopwords in the response ("lightning", "flash", and "sky") would be compared with the non-stopwords in the scenario ("light" and "darkness"). However, previous research has found that when using similarity to measure originality, at least with LSA, it is better to compare the potentially creative responses to non-creative responses (Forster & Dunbar, 2009). Specifically, Forster and Dunbar gathered uncreative data from a separate sample of participants; these participants were asked to generate common uses of objects from the alternate uses task, instead of uncommon uses. The similarity between the common and uncommon uses produced the greatest correlation with human judgments of creativity.

Therefore, in addition to the comparison with the scenario, we included a comparison of each response to common responses. In place of a data set of standard responses, we used the most common five words in the responses to the two scenarios. We excluded stop words and the words from the scenario (i.e., we excluded "light" and "darkness" from the first scenario and "cloth" and "breeze" from the second). In addition to the common words point of comparison, we also tested the correlation between the similarity of the responses to the words in the scenario.

Table 3 shows the correlation between the four different methods of representing phrases by the two different points of comparison with the mean external rating of each response. With one exception (LSA for the cloth scenario using the scenario point of comparison), all of the correlations were significant (p¡.001). Note that for this and subsequent tables, only significant correlations are shown. The largest correlations with the external raters was with the avePMI method of representing phrases using the common words point of comparison (-.542 for the light scenario and -.392 for the cloth scenario). In general (for 7 of the 8 cases), the common word point of comparison leads to larger correlations with the external raters than did the scenario point of comparison, replicating what was found by Forster and Dunbar (2009).

Table 3: Correlation between external rater judgments of originality and automated methods.

	Measure	Light	Cloth
Scenario	avePMI	-.424	-.184
	prodPMI	-.357	-.322
	w1PMI	-.382	-.272
	LSA	-.362	
Common Words	avePMI	**-.542**	**-.392**
	prodPMI	-.502	-.261
	w1PMI	-.387	-.273
	LSA	-.370	-.244
Other	Elaboration	.104	
	Ext. Raters	.711	.679
	Self-Ratings	.380	.329

Table 3 includes three other sets of results. The first is the correlations between the Elaboration (word count) metric and ratings of similarity. This measure correlates weakly with originality for the light scenario (.104, p=.025) and not significantly for the cloth scenario. The second is the average correlation among external raters. This was calculated by correlating each of the six external raters with the average of the other five for all responses. On average, external raters were correlated with each other at .711 for the light scenario and .679 for the cloth scenario. This could be seen as an upper limit of the degree to which the automated methods correlate with the external raters.

The final method reported in the table was the correlation between the external ratings and the self-ratings. This correlation was much lower than that of the external raters to each other, perhaps not surprisingly. What is of interest is that the avePMI method was more strongly correlated with external raters than were self-ratings. That is, the best of the automated scoring methods outperformed the human generated, self-ratings, when using human external raters as the gold standard.

One additional result shown in Table 3 worth noting is that of the eleven methods of predicting external ratings all of them performed worse for the cloth scenario than for the light

scenario. This included not only the automated methods but also the external and self-ratings. Also, only the avePMI with the common words point of comparison was able to perform better than the self-ratings for the cloth scenario.

Scoring Individual Participant Originality

Often assessments of creativity are used to assess individual differences. It is therefore important to determine how well the automated methods do scoring not just individual responses but also individual participants. To test this, we generated an average similarity score for each participant for each scenario and scoring method. Also, as this was a measure that included multiple responses per participant, we were able to apply the Fluency metric (the count of number of responses per participant per scenario).

As shown in Table 4, the correlations are generally at least slightly smaller when scoring by participant relative to scoring by response. Again only significant correlations are shown (p¡.05). Perhaps the most striking difference is for the automatic scoring methods when using the scenario point of comparison, as none of them remained significant predictors of the external ratings for the cloth scenario. Consistent with the individual response results, the common-words point of comparison leads to stronger correlations in all but one case (LSA). Overall, the avePMI representation performed the best, but was effectively tied (-.341, -.340, -.339 for the LSA, Fluency, and avePMI scoring methods, respectively) in predicting the cloth scenario with both LSA and the Fluency metric. The Fluency metric performed surprisingly well for both scenarios. It was not expected that simply counting the number of responses per participant would relate so strongly with the average originality rating of the responses.

Table 4: Correlation between external rater judgments and automated methods by participant.

	Measure	Light	Cloth
Scenario	avePMI	-.312	
	prodPMI	-.345	
	w1PMI	-.416	
	LSA	-.335	
Common Words	avePMI	**-.571**	-.339
	prodPMI	-.391	-.309
	w1PMI	-.483	
	LSA	-.311	**-.341**
Other	Elaboration		
	Fluency	.464	.340
	Ext. Raters	.647	.726
	Self-Ratings		.313

The avePMI method correlates more strongly with average participant originality than the self-ratings do. In addition, both LSA and the Fluency performed better than the self-ratings when participant scores were averaged. This combined with the fact that there was not a significant correlation between self-ratings and the external raters for the light scenario, indicates that self-ratings are not the best indicators of originality.

Discussion

The goal of the present study was to test the ability of similarity of semantic representations to predict human judgments of originality. To this end, we compared a number of different factors: type of semantic space, methods of comparison, and methods of phrase representation. We also compared the performance of semantically informed automated scoring methods against automated methods blind to semantic information.

Similarity and Originality

We found that similarity can be used to predict originality. Specifically, the similarity of a response to common response words to a scenario were negatively correlated with human judged originality of the response. This fits well with our previous work on flexibility, which found that similarity within a participant's set of responses was negatively correlated with the flexibility of the participant's responses (Blok et al., 2011). Therefore, a next step would be to test if combining these two aspects of creativity (originality and flexibility) would provide a predictor of human judgments of creativity than the two separately.

LSA vs. PMI

Comparing the semantic spaces, we found that PMI matched or exceeded the performance of LSA. These results provide additional support for larger data trumping smarter algorithms (Recchia & Jones, 2009). Previous results indicate that PMI is unable to outperform LSA when given the same corpus. It is the ability of PMI to make use of larger copora (i.e., more data) that allows the model to outperform LSA in general and specifically, in the present study. The corpus used for LSA was carefully created TASA corpus which is rather small compared to the entirety of Wikipedia, the corpus used by our PMI space.

Point of Comparison

The point of comparison, what participant responses were compared to for the originality scoring, had a large and consistent impact on performance. The better point of comparison for judging similarity was not the scenario or problem participants were given but instead uncreative responses (Forster & Dunbar, 2009). This was approximated in the present study by the most common words from the responses. This result was found both when using LSA and PMI.

PMI phrase representation

Unlike LSA, there is not an accepted standard for how to represent phrases within PMI. Of the three methods tested here, the average PMI between the words of the two phrases

worked the best for both predicting similarity judgments between phrases and for predicting originality judgments. The product of the pairwise comparisons performed well for certain comparison types, for example, for predicting similarity between two compound-noun phrases, but did not perform consistently. Furthermore, prodPMI always produced correlations at least slightly less than the avePMI.

The first-word method (w1PMI), a method that only used the PMI between the first words of the two phrases and ignoring all the other words, performed surprisingly well. While it was never the strongest method, it was also rarely the weakest. This fact suggests that either the entirety of the phrase was not that important for the present datasets or that the methods tested in the present research to represent the phrases were unable to capture much of the added meaning from the phrase. The latter possibility appears more likely.

Fluency and Elaboration

The two semantically uninformed automated scoring methods could be considered straw man metrics. Indeed, the Elaboration metric, the count of the number of words in each response, did quite poorly relative to the other measures. However, Fluency, which was only applicable for predicting originality at the participant level, did remarkably well. It was somewhat less accurate than the best PMI method tested for one scenario, but performed as well or better than LSA on the two scenarios. This result, combined with the relative ease of implementing the metric (i.e., tallying the number of responses per participant), suggests that this might be a worthwhile first-pass or heuristic method of scoring originality.

Conclusion

The present results suggest that similarity in semantic spaces can be used to predict originality in participant responses. In particular, the average pairwise PMI between a response and the most common responses to a scenario, performed the best at predicting originality. However, there are a number of caveats to these conclusions. First, the analysis was only conducted on two scenarios and the ability of the automated methods to score these two scenarios varied noticeably. Second, when scoring individuals, not individual responses, the Fluency metric, a tally of the number of responses per participant, did a surprisingly good job predicting originality. This combined with the success of the first-word only representation of phrases indicates that there is much room for improving phrase representations with semantic spaces.

Acknowledgments

This research was supported in part by the University of Maryland Center for Advanced Study of Language with funding from the Department of Defense.

References

Blacoe, W., & Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 joint conference on empirical methods in natual language processing and computational natural language learning* (pp. 546–556). Stroudsburg, PA: Association for Computational Linguistics.

Blok, S., Harbison, J. I., Haarmann, H. J., Bloodgood, M., & Berens, M. (2011). *Determining the feasibility of automatically scoring parts of the ardtt* (Tech. Rep. Nos. Part B of TTO 3503 Improving assessment of analyst-relevant divergent thinking: test validation, automated scoring, and brain signature). College Park, MD: University of Maryland, Center for Advanced Study of Language.

Budiu, R., Royer, C., & Pirolli, P. (2007). Modeling information scent: A comparison of LSA, PMI, and GLSA similarity measures on common tests and corpora. In *Proceedings of the 8th annual conference of recherche d'information assistee par ordinateur (RIAO)* (pp. 314–332). Pittsburgh, PA: Centre des Hautes Etudes Internationales d'Informatique Documentaire.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, R., Solan, E., Wolfman, G., et al. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*, 116–131.

Forster, E. A., & Dunbar, K. N. (2009). Creativity evaluation through latent semantic analysis. In *Proceedings of the 31st annual meeting of the cognitive science society* (pp. 602–607). Amsterdam, The Netherlands: Cognitive Science Society, Inc.

Fu, W.-T., & Pirolli, P. (2007). SNIF-ACT: a cognitive model of navigation on the world wide web. *Human-Computer Interaction*, *22*, 355–412.

Haarmann, H. J., O'Rourke, P., George, T., Smaliy, A., Grunewald, K., Dien, J., et al. (2011). *Improving assessment of analyst-relevant divergent thinking: test validation, automated scoring, and brain signature* (Tech. Rep. No. TTO 3503). College Park, MD: University of Maryland, Center for Advanced Study of Language.

Kimball, D. R., Smith, T. A., & Kahana, M. J. (2007). The fSAM model of false recall. *Psychological Review*, *114*, 954–993.

Landauer, T. K., Foltz, P. W., & Laham, D. (1983). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, *34*, 1388–1429.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparising pointwise mutual information with latent semantic analysis. *Behavioral Research Methods*, *41*, 647–656.

Wang, H. C., Chang, C. Y., & Li, T. Y. (2008). Assessing creative problem-solving with automated text grading. *Computers & Education*, *51*, 1450–1466.

Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (Eds.). (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone.