Article

Educational and Psychological Measurement
2020, Vol. 80(2) 399–414

DOI: 10.1177/0013164419860575
journals.sagepub.com/home/epm
SAGE

Using Latent Semantic Analysis to Score Short Answer Constructed Responses: Automated Scoring of the Consequences Test

Noelle LaVoie[1], James Parker[1], Peter J. Legree[2], Sharon Ardison[2] and Robert N. Kilcullen[2]

Abstract

Automated scoring based on Latent Semantic Analysis (LSA) has been successfully used to score essays and constrained short answer responses. Scoring tests that capture open-ended, short answer responses poses some challenges for machine learning approaches. We used LSA techniques to score short answer responses to the Consequences Test, a measure of creativity and divergent thinking that encourages a wide range of potential responses. Analyses demonstrated that the LSA scores were highly correlated with conventional Consequence Test scores, reaching a correlation of .94 with human raters and were moderately correlated with performance criteria. This approach to scoring short answer constructed responses solves many practical problems including the time for humans to rate open-ended responses and the difficulty in achieving reliable scoring.

Keywords

automated scoring, Latent Semantic Analysis, short answer scoring, creativity, constructed responses, LSA

[1]Parallel Consulting, Petaluma, CA, USA
[2]U.S. Army Research Institute for the Behavioral and Social Sciences, Fort Belvoir, VA, USA

Corresponding Author:
Noelle LaVoie, Parallel Consulting, 10 Arlene Court, Petaluma, CA 94952, USA.
Email: lavoie@parallel-consulting.com

We evaluated the effectiveness of using Latent Semantic Analysis (LSA) to score open-ended short answer responses. LSA-based automated scoring is routinely used for large-scale scoring of essays responses on high- and low-stakes exams (Shermis, 2014; Zhang, 2013), and LSA has been used for short answer scoring (Streeter, Bernstein, Foltz, & DeLand, 2011). To improve the accuracy of short answer scoring, LSA has been limited to test items with limited potential responses rather than items with unconstrained potential responses. To automatically score tests that measure constructs such as creativity, a new approach to automated scoring is required—one that can accurately score short responses with minimally constrained answers. We used LSA techniques to score short answer responses to the Consequences Test, a measure of creativity and divergent thinking, which encourages a wide range of potential responses.

Latent Semantic Analysis

We chose to use LSA as the machine learning technique for this application because it has several important properties that make it suitable for automated scoring in general and scoring of creativity tests that require examinees to provide open-ended short answers. LSA is both a machine learning technology and theory of knowledge representation (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). LSA uses a fully automatic mathematical technique to extract and infer meaning relations from the contextual usage of words in large collections of natural discourse. LSA simulates important practical aspects of human meaning to a very useful level of approximation. And it does so by using a robust, domain-independent process that allows it to effectively perform tasks that, when performed by a human, depend on understanding the meaning of textual language.

LSA takes as input large quantities of raw text parsed into words and separated into meaningful passages such as sentences or paragraphs. Although the technique is based on the statistics of how words are used in ordinary language, its analysis is much deeper and more powerful than simple frequency, co-occurrence, or keyword counting and matching techniques. LSA infers the relations between words and documents by machine analyzing large collections of text in two distinct steps.

First, a background semantic space is built. The semantic space is similar to a large training set that provides LSA with a full context for evaluating potential responses. The background space is developed from a very large collection of text and must contain a minimum of 100,000 paragraphs of text (Landauer, McNamara, Dennis, & Kintsch, 2007). Materials included in the semantic space are chosen to ensure that all vocabulary terms that may be found in the analyzed text are included. The quality of the semantic space is a major consideration for the success of using LSA. The semantic space is not a repository of essays, rather it is a background space that allows the system to infer general semantic relationships (Martin & Berry, 2010). The output of this analysis is a several hundred dimensional semantic space in which every word

and every document is represented by a vector of real numbers—one vector for each dimension.

Once the semantic space is created, the second step is to project new segments of text (e.g., short answer responses) into the semantic space for analysis. These text segments are represented as vectors in the space. These vectors can be compared with one another using metrics that capture the relative semantic similarity among the texts such as the cosine of the angles between the vectors representing two responses. Among other possibilities, an LSA similarity score can then be computed for the new text by quantifying its similarity to the previously scored texts.

LSA is not a traditional natural language processing or artificial intelligence program; it uses no humanly constructed ontologies, dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies (Landauer et al., 1998). LSA provides an important conceptual advance on ''key word search'' algorithms because this technology provides the basis to assess the semantic similarity of the content heavy nouns and verbs that provide most of the meaning in textual passages.

To understand this capability, consider sample short answer responses such as ''no more dinner'' and ''skip breakfast.'' From a key word search perspective, these two phrases do not contain any common words and would appear independent. However, these two phrases contain terms that are semantically similar on multiple dimensions: dinner and breakfast are both meals, no and skip both imply an absence, and more generally, these phrases may have similar meaning within a common paragraph. LSA algorithms can be used to assess the similarity of the separate terms that appear in these two phrases and thereby quantify the semantic similarity of these phrases. So, unlike a key word search algorithm that would infer no overlap between these phrases, LSA would judge these two phrases to be semantically similar as represented by the cosine between the vectors for each phrase.

It is also worth noting that newer techniques, including Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) are sometimes used instead of LSA to examine topics in text. Empirical comparisons among these techniques shows that the best performance varies considerably by application. LSA tends to be superior for tasks that require behavior that mimics human assessments of the similarity of texts such as recommending movies (Bergamaschi & Po, 2015). LSA also allows the direct comparison of similarity among documents. LDA is more precise in identifying individual emails as spam than either LSA or PLSA, while PLSA has been shown to outperform LSA in automatically indexing documents, for example, adding specific tags to a document (Hoffman, 2001). For tasks such as classifying texts, LSA and LDA perform similarly, with a slight advantage for LSA (Anaya, 2011; Cvitanic, Lee, Song, Fu, & Rosen, 2016). For the current task, judging the creativity of responses, we chose to use LSA because human-like judgments are required for this application.

LSA Assessments

Essay Analyses. LSA scoring algorithms can be used to quantify individual differences in the extent to which respondents have used the correct words within an essay

on a specific topic (Landauer, Laham, & Foltz, 2003), and analyses have shown that LSA is useful in assessing the quality of essays that have been written by respondents on specific topics (Landauer, Laham, & Foltz, 2000). LSA technologies have been adapted to support automated essay scoring in educational settings to provide writing instruction (Streeter, Berstein, Foltz, & DeLand, 2011), as well as for high-stakes writing assessments including the SAT, GRE, and GMAT (Shermis, 2014; Zhang, 2013) and low-stakes writing assessments such as evaluating military leadership and medical diagnostic reasoning (Landauer et al., 2000; LaVoie, Cianciolo, & Martin, 2015; LaVoie et al., 2010). Analyses also demonstrate that LSA-generated scores often have high agreement with subject matter experts (SMEs), that is on par with agreements between SMEs (Landauer et al., 2000, 2003; Shermis, 2014; Shermis, Burstein, Higgins, & Zechner, 2010).

Short Constructed Responses. LSA has been used to score short constructed response items. However, it has been more challenging to reach high agreement with SMEs for these applications because there is significantly less text to analyze. For this reason, much of the work on scoring short constructed responses has been focused on constrained test items that assess content knowledge, such as questions that require brief explanations of science concepts (Liu et al., 2014). Streeter and her colleagues (2011) report that in more than 5 years of scoring short answers as many as 50% of the short answer items from a state science test cannot be scored accurately enough for high stakes testing. Likewise, Liu and associates (2014) found that short answer scoring was more difficult when rubrics had to accommodate increased complexity of ideas. Finally, Kersting, Sherin, and Stigler's (2014) automated scoring of a short answer assessment designed to measure teacher knowledge yielded moderate to high correlations with human scores. Thus, successfully scoring unconstrained short answer responses will require going beyond current automated scoring capabilities.

Measuring Creativity

Accurate measurement of creativity is important because it can improve predictions of leadership and academic performance (Stemler, Sternberg, Grigorenko, Jarvin, & Sharpes, 2009; Sternberg, 2015; Zaccaro, Mumford, Connelly, Marks, & Gilbert, 2000), but the very nature of creativity makes it difficult. One of the most commonly used measures of creativity and divergent thinking is Guilford's Consequences Test (Christensen, Merrifield, & Guilford, 1953). The Consequences Test items require participants to list, in 2 minutes, as many consequences as possible for an unusual situation framed as ''What would be the result if . . .'' (Guilford & Guilford, 1980). Each response consists of briefly described consequences, typically only a few words long. For example, a Consequences Test scenario might ask respondents, ''What would be the result if people no longer needed to eat?'' An individual might respond to this question by listing the following responses: no more cooking shows, no more dinner, skip breakfast.

Consequences Test scores are intended to quantify the creativity of individual responses based on the subjective judgment of SMEs (Guilford & Guilford, 1980). However, SMEs must also ensure that individual scores are not inflated as a result of double counting responses that are conceptually duplicative (e.g., ''no more dinner'' vs. ''skip breakfast''). Due to the subjective nature of the scoring process, maintaining consistent scoring standards across raters and over time has been challenging. Therefore, scoring the Consequences Test requires that several SMEs independently assess each protocol and then meet as a team to resolve differences in opinion. Due to the complexity of this process, scoring the Consequences Test has been both time-consuming and labor intensive.

Despite the difficulty in scoring the Consequences Test, performance on this scale has been shown to predict important aspects of career continuance and performance of senior leaders in the U.S. Army (Mumford, Marks, Connelly, Zaccaro, & Johnson, 1998; Zaccaro, Mumford, Connelly, Marks, & Gilbert, 2000). However, reliance on humans to score the Consequences Test makes it impractical to administer on a large scale. To address the limitation of human scoring, we used LSA to automatically score responses to the Consequences Test.

Automated Scoring of the Consequences Test

As discussed, short answers and particularly unconstrained, creative short answers pose significant challenges for automated scoring, suggesting that it may be difficult to develop automated scoring for this test. The Consequences Test represents an unusual application for automated scoring because we need to ensure that the technology can identify both creative responses and duplicative responses. On the other hand, responses to the Consequences Test usually consist of content-heavy nouns and verbs, to which LSA is most sensitive. In addition, multiple responses are given to each Consequences Test item so that the total number of responses available for developing automated scoring algorithms is fairly high. This project addressed the following questions:

- Can an LSA-based automated scoring model for the Consequences Test reach the same level of agreement as two trained SMEs?
- Will LSA scores of the Consequences Test have the same patterns of predictive and concurrent validity as human scores with a set of outcome and concurrent measures?

Method

Participants

A sample of 1,863 participants provided responses to the Consequences Test. This sample was drawn from a larger sample of 5,191 Reserve Officers' Training Corps cadets who participated in a U.S. Army training exercise during the summer of 2013.

From a demographic perspective, the sample was approximately 78% male, 82% Caucasian, 11% African American, 7% Asian, 2% American Indian or Alaskan Native, and 1% Native Hawaiian or other Pacific Islander. In addition, 12% of the sample identified their ethnicity as Hispanic. Approximately 2% of the sample did not identify their race or ethnicity.

Measures

Consequences Test. Five items from the Consequences Test were administered. Each item required participants to list as many consequences of an unusual situation framed as ''What would be the result if . . .'' (Guilford & Guilford, 1980). The Consequences Test was scored in three ways: trained human ratings, automated scores, and word count.

Concurrent Measures. Concurrent predictors included scales derived from the Cadet Background and Experiences Form (CBEF). The CBEF is a multiple-choice questionnaire assessing past behaviors and experiences that are related to officer performance and retention (Putka, 2009). All items within the CBEF scales are measured using a 5-point Likert-type scale. The following CBEF scales were included in the analyses: Achievement Orientation (ACH), Army Identification (AI), Fitness Motivation (FM), Peer Leadership (PLEAD), Stress Tolerance (ST), and Lie (LIE). Two additional measures were also analyzed: the Leader Knowledge Test Characteristics (LKT Char) and Skills (LKT Skills) in which respondents rate the importance of 30 characteristics and 30 skills associated with leadership.

Outcome Measures. Following LDAC completion, outcome data were collected to validate the Consequences Test scores: Cadet Grade Point Average (GPA), and Cadet Order of Merit Listing (OMS). OMS combines performance outcomes including college GPA, training cadre ratings, military science course grades, and cadet physical fitness scores.

Procedure

The process of developing an effective LSA scoring model requires reliable human scores for training the automated scoring system. Thus, the first step in developing a scoring model is to have trained SMEs score all the responses. These scores are also used to evaluate the model's consistency with human ratings using a hold-out ''test'' sample.

Human Scoring of the Consequences Test. Our human scoring approach started by separating the scoring categories into two orthogonal scales because we intended to use two separate automated scoring techniques, one for remote and obvious responses and one for duplicate responses. The first scale was content, which included remote (score = 2), obvious (score = 1), and irrelevant responses (score = 0). The second

scale categorized responses as either unique (score = 1) or duplicate (score = 0). A detailed scoring rubric was created for each scale, based on Guilford and Guilford (1980), in conjunction with a set of anchors, or sample responses, that exemplified each category. The initial anchors were drawn from examples provided in the Consequences Test scoring booklet (Guilford & Guilford, 1980). This initial set was augmented during the process of training the raters to achieve consistent scoring.

Responses from 420 participants had previously been scored by a team of U.S. Army personnel. These responses were used to train our two SMEs to use the rubrics and anchors. The two raters were trained to score the responses in a multistep process designed to maintain high agreement between the raters and consistency with the U.S. Army–scored sample. The rating process required both SMEs to independently rate a set of 50 responses, compare their ratings to identify any discrepancies, and discuss the discrepancies to reach a consensus on the correct category. Then, the raters compared their scores with the sample scored by the U.S. Army and identified any inconsistencies. These inconsistencies were then discussed to determine if the rubric or anchor sets needed to be updated. This training process was repeated until all 420 responses with U.S. Army ratings were scored.

Once the SMEs were trained, they scored the responses from the remaining 1,443 participants on all five Consequences Test items. Because each participant could list as many or few responses to an item as they chose, the number of responses varied by item from 5,596 to 7,300. This rating process involved scoring responses in batches of 100. After each batch of responses were individually scored, raters compared their ratings to identify any discrepancies and discuss the discrepancies to reach a consensus on the correct category for each response. When raters could not reach consensus, a third rater was used. Raters were asked to review the scoring rubric and anchors before beginning each new batch to help maintain consistency over time. Raters scored all responses on the content scale first, then scored all responses on the duplicate scale.

Human Scoring Agreement. Agreement between the raters was calculated for the total Consequences Test score using Pearson correlations, exact agreement (percentage of responses where both raters had the same score), and adjacent agreement (percentage of responses where raters were within 1 score point). Agreement between the two raters was high across the 1,374 cases with a correlation of $r = .98$, exact agreement of 26%, and adjacent agreement of 64%. The total human consensus Consequences Test scores ranged from 4 to 72.

Automated Scoring Process. The automated scoring system relies on LSA and requires two components: a background semantic space and a set of scoring algorithms. The goal of the automated scoring system was to correctly categorize each response on the content scale, assigning a score of 0, 1, or 2 (irrelevant, obvious, or remote), and on the duplicate scale, assigning a score of 0 or 1 (duplicate or unique). The response scores were then summed for each Consequences Test item, and the item scores were

summed to create total scores for each participant across their responses on the five test items. Separate automated scoring processes were created for each of the two scales. The process of developing each component is described in turn.

Background semantic space. The semantic space is similar to a large training set that provides LSA with a full context for evaluating responses. The background space is developed from a very large collection of text and must contain a minimum of 100,000 paragraphs of text (Landauer et al., 2007). A background semantic space is created by automatically analyzing a large body of text to extract latent knowledge of a domain and can be used to measure similarity of meaning between multiple texts. We constructed a semantic space with 100,000 paragraphs of general written language drawn from the Reuters News corpora. This corpora was chosen as the basis for the space because the Consequences Test does not cover any particular domain in depth, but rather encourages responses on a wide range of topics. We then evaluated the semantic space for any gaps in content by identifying any words present in the Consequences Test responses but missing from the semantic space. There were 126 unique missing terms. These missing terms guided the selection of an additional 18,000 paragraphs of text selected from Wikipedia to fill in the content gaps. The final semantic space contained 118,000 documents, and the remaining missing terms were primarily misspelled words present in the Consequences Test responses.

Content scoring algorithm. The content scoring used a combination of two automated scoring techniques: content clustering and anchor mapping. For clustering, the responses were first separated into groups by score (0, 1, and 2) and then projected into the LSA background space. Using a clustering algorithm, the responses were automatically grouped by topic so that responses with similar meanings were grouped together within each score point. This process created hundreds of clusters in the LSA space.

The second process, anchor mapping, relied on the anchors (sample responses) that the raters used to score the responses. The anchors are representative sample responses at each score point drawn from the Consequences Test scoring booklet (Guilford & Guilford, 1980), and the first 420 responses that the raters were trained on. On average, there were 66 remote (score = 2) anchors, 68 obvious (score = 1) anchors, and 3 irrelevant (score = 0) anchors for each item. These anchors were not included in the training or testing data. The anchors were projected into the LSA space.

After the topic clusters were formed and the anchors were projected into the LSA space, there was a combination of topic clusters and anchors present in the space representing each score point. To score a new response, the response was projected into the LSA space alongside the clusters and anchors. Then, the semantic similarity between the response and every cluster and anchor was calculated as a cosine between the response and each cluster and anchor. The nine clusters and anchors with the highest cosine to the response were identified as the response's nearest neighbors in the LSA space and were the clusters and anchors that were most similar in meaning to the response being scored. Once the nearest neighbors were identified, the modal, or most frequently occurring score, was assigned as the score for the response. For example, if the clusters and anchors identified as nearest neighbors had

the following scores: 0, 1, 1, 1, 2, 1, 1, 2, 1, 1, the response received a score of 1. This was the process used to create the automatic content score for each individual response.

Duplicate scoring algorithm. Duplicates were defined as responses that repeated a previous response made by the participant to the item or repeated an example included as part of the test item. Duplicate scoring was handled by comparing each response with the example responses provided in the Consequences Test questions and to the participants' other responses to the item. These responses were projected into the LSA background space and cosines were calculated between the response to be scored, the example responses, and the participant's other responses. If any responses exceeded a cosine of .70 with either the example responses or the participants' other responses, they were scored as a duplicate. For example, a response of ''no more dinner'' would be compared with the participant's other responses, including ''no more cooking shows'' and ''skip breakfast.'' Because ''no more dinner'' has a very similar meaning to ''skip breakfast,'' the cosine between the two would exceed the threshold and ''skip breakfast'' would be scored as a duplicate. This process was used for automatically identifying all duplicate responses.

The final step was to combine the content and duplicate scores. The content and duplicate scores were multiplied for each response to create a score for each response. The scores for each response were then summed to calculate item scores for each of the five items. Finally, the item scores were summed to create total Consequences Test scores for each participant.

Word count of the Consequences Test. We also computed word count as an alternate scoring option for the Consequences Test. Word count provides a straightforward and easy–to-calculate measure of verbal fluency. The total length of responses to all five items on the Consequences Test varied from 24 words to 265 words with a mean of 112.6 words.

Results

The automated scoring was evaluated in two ways. First, the consistency of LSA Consequences Test scores was assessed by comparing them with SME Consequences Test ratings. Second, the validity of the automated scores was calculated by examining the correlations between the automated scores, the human scores, and performance measures that were collected from participants at the same time as the Consequences Test responses. Finally, we consider the role of response length in the performance of the automated scoring model and whether word count would be a suitable alternative to automated scoring.

Convergence of LSA and SME Consequences Test Scores

The association between the LSA scores and the SME Consequences Test scores was evaluated across four separate holdout sets to ensure generalizability. These were

Table 1. SME and LSA Score Agreement for Consequences Test Total Scores.

Holdout set	Comparison	*n*	Pearson *r*	Exact agree (%)	Adjacent agree (%)
A	Rater 1–Rater 2	677	.98**	25	64
A	Raters–Auto	677	.94**	13	36
B	Rater 1–Rater 2	689	.98**	26	64
B	Raters–Auto	689	.95**	14	40
C	Rater 1–Rater 2	691	.98**	25	63
C	Raters–Auto	691	.94**	13	34
D	Rater 1–Rater 2	690	.98**	25	65
D	Raters–Auto	690	.94**	11	36
Average	Rater 1–Rater 2	687	**.980**	25.3	64.0
Average	Raters–Auto	687	**.943**	12.8	36.5

Note. Auto = automated score; The rows marked average provide the average statistics across the 4 hold out sets to make it easier for readers to see how well LSA scoring agreed with the SME scores.
**$p < .001$ (two-tailed).

created by randomly selecting one half of the data set and using it for training leaving the remaining half for testing. For each set, reliability and agreement were calculated by comparing the automated score with the human consensus score. These were compared with the agreement between the two human ratings prior to reaching consensus.

The results are shown in Table 1. The performance of the automated scoring approach was comparable across all four holdout sets indicating that the automated scoring approach is generalizable across all four training and testing sets. Averaged across the four sets, the correlation between the automated scores and human-rated scores was $r = .94$, $n = 687$, $p < .001$, which approached the correlation between the two human ratings of $r = .98$, $n = 687$, $p < .001$, indicating a very high level of convergence between automated scoring and human scoring.

Concurrent Validity of Automated Scores

The concurrent validity of the automated scores was calculated using the set of concurrent measures including the two outcome criteria as well as eight predictors. Because half of the data were used to train the automated scoring system, only the holdout data were included in this analysis. The automated scores and human consensus scores of the Consequences Test have identical correlations with OMS, $r = .14$, $n = 471$, $p < .01$, and very similar correlations with GPA, automated scores $r = .09$, $n = 496$, $p = .04$, and human consensus scores $r = .10$, $n = 496$, $p = .02$. This result indicates that the LSA scores provide parallel results with the SME scores for the purpose of predicting the primary outcome criteria (i.e., GPA and OMS).

Finally, we observed that the LSA and SME Consequences Test scores had a very similar pattern of correlations with the temperament predictors that are listed in Table 2. In fact, as reported in Table 3, the correlation between the two sets of coefficients

was very high, $r = .96$, $p < .001$. This last observation suggests a high degree of equivalence between the LSA and SME scores.

The predictive value of the LSA Consequences Test scores was evaluated by comparing a model made by regressing OMS on the available concurrent measures with a model that included the same set of measures plus the automated Consequences Test scores, $F(2, 191) = 2.30$, $p = .13$. The Consequences Test scores did not add incremental validity to the prediction of OMS over and above the other variables.

Comparison of LSA With Word Count

Automated scoring models are known to be sensitive to response length (Attali, 2013) so we calculated the partial correlation between automated scores and human consensus scores controlling for the number of words in the response, $r = .88$, $n = 675$, $p < .001$. This estimate demonstrates that the majority of the shared variance between automated scores and human scores is independent of length. We also compared the performance of the automated scoring model with word count. Word count had lower correlations with the two outcome measures than the automated scores or human consensus scores: OMS ($r = .11$, $n = 471$, $p = .02$) and GPA ($r = .08$, $n = 471$, $p = .06$). Word count also had correlations that were equivalent or lower than the human consensus scores and automated scores with FM, LIE, LKT Skills, PLEAD, and ST. The correlations between word count and ACH, AI, and LKT Char were somewhat higher than the correlations with the human consensus scores or automated scores (see Table 2). The automated scoring model performs better than word count alone.

Discussion

In this article, we have shown that LSA is appropriate for large-scale scoring of the Consequences Test, a short answer constructed response test of creativity. Using topic clusters and anchor items to train the scoring model, we were able to come very close to SMEs' level of agreement, reaching a very high correlation of .94 with the human ratings, despite the challenges involved in automatically scoring open-ended short answer responses. The automated LSA scores also achieved excellent convergence with SME ratings indicating that an automated scoring model may be used to effectively score a measure of creativity and divergent thinking in lieu of the cumbersome human scoring process.

The automated scores showed very similar patterns of correlations with several measures demonstrating the validity of the scoring approach against cadet outcome variables and concurrent measures. However, the Consequences Test did not improve predictions of cadet outcomes over and above a set of other measures. Nevertheless, we expect to find a higher incremental validity of the Consequences Test in a sample with higher command groups of more advanced age (e.g., Officers instead of Cadets) as suggested by both Spearman's Law of Diminishing Returns (SLODR), the finding

Table 2. Concurrent Validity of Automated Scores, Pearson Correlations.

	ACH	AI	FM	GPA	LIE	LKT Char	LKT Skills	OMS	PLEAD	ST	LSA	SME	WC
ACH	1.0	.14*	.20*	.32*	.12	.05	.03	.25*	.35*	.08	.10	.08	.17*
AI		1.0	.36*	−.25*	.12*	.08	.02	−.13	.26*	.15*	.03	.02	.04
FM			1.0	−.08	.12	.06	.05	.18*	.35*	.35*	.11	.07	.05
GPA				1.0	.01	.05	.17*	.82*	.00	.03	.09*	.10*	.08
LIE					1.0	−.16*	−.16*	.05	.09	.18*	−.13*	−.13*	−.12*
LKT Char						1.0	.44*	.03	.00	.15*	.06	.07	.08*
LKT Skills							1.0	.17*	.08	.15*	.05	.08*	.05
OMS								1.0	.10	.13	.14*	.14*	.11*
PLEAD									1.0	.14*	.19*	.14*	.07
ST										1.0	.04	.02	.01
LSA											1.0	.94*	.69*

Note. ACH = Achievement Orientation; AI = Army Identification; FM = Fitness Motivation; GPA = Grade Point Average; LIE = Lie; LKT Char = Leader Knowledge Test Characteristics; LKT Skills = Leader Knowledge Test Skills; OMS = Order of Merit Listing; PLEAD = Peer Leadership; ST = Stress Tolerance; LSA = Latent Semantic Analysis; SME = subject matter expert. SME indicates rater consensus score and LSA indicates automated score. *n*(ACH), *n*(AI), *n*(FM), *n*(LIE), *n*(PLEAD), *n*(ST) = 275, *n*(GPA) = 498, *n*(LKT Char), *n*(LKT Skills) = 633, *n*(OMS) = 473.

*$p < .05$ (two-tailed).

Table 3. Pearson Correlations Between Regression Coefficients.

	LSA	SME	WC
LSA	1.0	.963**	.784*
SME		1.0	.834**
WC			1.0

Note. n = 8., LSA = Latent Semantic Analysis; SME = subject matter expert; WC = word count.
*$p < .05$ (two-tailed). **$p < .01$ (two-tailed).

that cognitive tests are less correlated in higher ability groups (Blum & Holling, 2017), and the age differentiation hypothesis, which suggests that ability tests are less correlated in older groups as cognitive ability becomes more differentiated with age (McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002). Future work with samples drawn from older, more senior officers, will explore this prediction.

The automated Consequences Test scores provide more information than just fluency, and the agreement between Consequences Test scores and human scores is based on more than simple word count. It is also interesting to note that the pattern of correlations in Table 3 shows higher correlations between word count and the SMEs than between word count and the LSA scores, which suggests that the automated scores may actually be less sensitive to word length than the human raters.

To successfully develop automated scoring of the open-ended short responses typical of the Consequences Test, several requirements were met. A large sample size was used for training the scoring model (Streeter et al., 2011). Trained SMEs provided ratings with a very high level of agreement, suitable for training an automated scoring model. And, a sufficient set of sample responses were used to anchor the human ratings and the scoring model.

Limitations

The utility of this approach to automated scoring is limited by the quality of the ratings used to train the automated scoring system. In this case, SMEs were first trained on a subset of items scored by U.S. Army SMEs. As a result, the quality of the U.S. Army SME ratings represents the upper limit of the quality of the automated scores. Improving the validity of the scores used to develop the automated scoring system could potentially improve the performance of the automated scores.

Word count, a simple measure of verbal fluency, was correlated with the Consequences Test automated scores. Overall patterns of correlations between word count and several measures and outcomes showed lower correlations than human ratings, with a few exceptions, suggesting that word count is not as effective a means of scoring the Consequences Test as LSA. The partial correlation between the automated scores and human rater scores, controlling for response length, was significant

indicating that word count is not responsible for all shared variance between the automated and human scores.

Conclusion

Automated scoring is a viable alternative to time-consuming hand scoring of the Consequences Test, suggesting that LSA-based scoring could be used to score other tests with open-ended short answer responses. This approach to scoring solves many practical problems, including the need to train SMEs, the time for humans to rate open-ended responses, and the difficulty in achieving reliable scoring.

Authors' Note

The views contained in this article should not be construed as an official U.S. Department of the Army or U.S. Department of Defense position, policy, or decision, unless so designated by other documentation. The opinions and findings are those of the authors and do not necessarily represent the views of the U.S. Army Research Institute for the Behavioral and Social Sciences.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by U.S. Army Research Institute Contract W911NF-13-C-0083 to Parallel Consulting.

ORCID iD

Noelle LaVoie https://orcid.org/0000-0002-7013-3568

References

Anaya, L. H. (2011). *Comparing latent dirichlet allocation and latent semantic analysis as classifiers* (Unpublished doctoral dissertation). University of North Texas, Denton, TX. Retrieved from https://digital.library.unt.edu/ark:/67531/metadc103284/m2/1/high_res_d/dissertation.pdf

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181-199). New York, NY: Routledge.

Bergamaschi, S., & Po, L. (2015, December). *Comparing LDA and LSA topic models for content-based movie recommendation systems*. Paper presented at the International Conference on Web Information Systems and Technologies, Lisbon, Portugal.

Blum, D., & Holling, H. (2017). Spearman's law of diminishing returns. A meta-analysis. *Intelligence*, *65*, 60-66.

Christensen, P. R., Merrifield, P. R., & Guilford, J. P. (1953). *Consequences Form A-1*. Beverly Hills, CA: Sheridan Supply.

Cvitanic, T., Lee, B., Song, H. I., Fu, K., & Rosen, D. (2016, October-November). *LDA v. LSA: A comparison of two computational text analysis tools for the functional categorization of patents*. Paper presented at the International Conference on Case Based Reasoning, Atlanta, GA.

Guilford, J. P., & Guilford, J. S. (1980). *Consequences: Manual of instructions and operations*. Orange, CA: Sheridan Psychological Services.

Hoffman, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*, 177-196.

Kersting, N. B., Sherin, B. L., & Stigler, J. W. (2014). Automated scoring of teachers' open-ended responses to video prompts: Bringing the Classroom-Video-Analysis assessment to scale. *Educational and Psychological Measurement*, *74*, 905-926.

Landauer, T. K., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259-284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The intelligent essay assessor. *IEEE Intelligent Systems*, *15*(5), 27-31.

Landauer, T. K., Laham, R. D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the intelligent essay assessor. *Assessment in Education*, *10*, 295-308.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum.

LaVoie, N., Cianciolo, A., & Martin, J. (2015, April). *Automated assessment of diagnostic skill*. Poster presented at the CGEA CGSA COSR conference, Columbus, OH.

LaVoie, N., Streeter, L., Lochbaum, K., Wroblewski, D., Boyce, L.A., Krupnick, C., & Psotka, J. (2010). Automating expertise in collaborative learning environments. *Journal of Asynchronous Learning Networks*, *14*(4), 97-119.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19-28.

Martin, D. I., & Berry M. W. (2010). Latent Semantic Indexing. In M. J. Bates & M.N. Maack (Eds.), *Encyclopedia of Library and Information Sciences* (pp. 3195-3204). New York, NY: Taylor & Francis.

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, *38*, 115-142.

Mumford, M. D., Marks, M. A., Connelly, M. S., Zaccaro, S. J., & Johnson, J. F. (1998). Domain-based scoring of divergent-thinking tests: Validation evidence in an occupational sample. *Creativity Research Journal*, *11*, 151-163.

Putka, D. J. (Ed.). (2009). *Initial development and validation of assessments for predicting disenrollment of four-year scholarship recipients from the Reserve Officer Training Corps* (Study Report 2009-06). Arlington, VA: U.S. Army Research Institute for the Behavioral Sciences.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53-76.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw, & N. S. Petersen (Eds.), *International encyclopedia of education* (3rd ed., pp. 75-80). Oxford, England: Elsevier.

Stemler, S. E., Sternberg, R. J., Grigorenko, E. L., Jarvin, L., & Sharpes, K. (2009). Using the theory of successful intelligence as a framework for developing assessments in AP physics. *Contemporary Educational Psychology*, *34*, 195-209.

Sternberg, R. J. (2015). Successful intelligence: A model for testing intelligence beyond IQ tests. *European Journal of Education and Psychology*, *8*, 76-84.

Streeter, L., Bernstein, J., Foltz, P., & DeLand, D. (2011). *Pearson's automated scoring of writing, speaking, and mathematics* (White Paper). Retrieved from http://kt.pearsonassessments.com/download/PearsonAutomatedScoring-WritingSpeakingMath-051911.pdf

Zaccaro, S. J., Mumford, M. D., Connelly, M. S., Marks, M. A., & Gilbert, J. A. (2000). Assessment of leader problem-solving capabilities. *Leadership Quarterly*, *11*, 37-64.

Zhang, M. (2013). *Contrasting automated and human scoring* (ETS Research Report No. RDC-21). Princeton, NJ: ETS. Retrieved from https://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf