Research Note 2022-05

Developing an Automated Scoring System to Support Soldier Assessments with the Consequences Test

Noelle LaVoie
James T. Parker
Parallel Consulting

Peter J. Legree
Mark C. Young
Robert N. Kilcullen
U.S. Army Research Institute

November 2021

United States Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

Department of the Army
Deputy Chief of Staff, G1

Authorized and approved:

MICHELLE L. ZBYLUT, Ph.D.
Director

Research accomplished under contract
for the Department of the Army by:

Parallel Consulting

Technical Review by:

Cristina Kirkendall, U.S. Army Research Institute

DISPOSITION

This Research Note has been submitted to the
Defense Technical Information Center (DTIC).

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from….to)
November 2021	Interim	August 2019 – September 2020

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Developing an Automated Scoring System to Support Soldier Assessments with the Consequences Test	W911NF-19-C-0097
	5b. PROGRAM ELEMENT NUMBER: 622785
6. AUTHORS: Noelle LaVoie, James T. Parker, Peter J. Legree, Mark C. Young, and Robert N. Kilcullen	5c. PROJECT NUMBER: A790
	5d. TASK NUMBER
	5e. WORK UNIT NUMBER: 1011
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES): Parallel Consulting, 10 Arlene Court, Petaluma, CA 94952	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES): U.S. Army Research Institute for the Behavioral and Social Sciences, 6000 6th Street (Bldg 1464/Mail Stop 5610), Fort Belvoir, VA 22060-5610	10. MONITOR ACRONYM: ARI
	11. MONITOR REPORT NUMBER: Research Note 2022-05

12. DISTRIBUTION/AVAILABILITY STATEMENT: Distribution Statement A.
Approved for public release, distribution unlimited.

13. SUPPLEMENTARY NOTES
ARI Research POCs: Dr. Mark C. Young and Dr. Peter J. Legree, Selection and Assignment Research Unit

14. ABSTRACT *(Maximum 200 words)*:

Measures of creativity, and particularly divergent thinking, such as the Consequences Test, have been shown to be good predictors of important aspects of career performance in the Army, including continuance and progression. Until recently, the requirement to have expert humans score the responses has made this test impractical to use on a large scale. Following the development of automated scoring models that can score these responses as well as expert humans, this test of creativity is being used operationally by the U.S. Army. The latest effort developed new scoring models for four additional test items, increasing the available number of test items in order to improve test security. We also developed new scoring tools to allow the Army to score these test items, along with five other test items whose scoring models were developed in an earlier project. The new scoring tools include an easy to use laptop version and a version intended for deployment on an Army server, allowing it to be integrated with other assessment programs.

15. SUBJECT TERMS
Consequences Test, Automated scoring, Latent Semantic Analysis, Automated scoring algorithms, Creativity, Army officer assessment

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17.ABSTRACT	18.THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited Unclassified	18	Tonia S. Heffner, Ph.D. (703) 545-4408

Research Note 2022-05

Developing an Automated Scoring System to Support Soldier Assessments with the Consequences Test

Noelle LaVoie
James T. Parker
Parallel Consulting

Peter J. Legree
Mark C. Young
Robert N. Kilcullen
U.S. Army Research Institute

Selection and Assignment Research Unit
Tonia S. Heffner, Chief

November 2021

DEVELOPING AN AUTOMATED SCORING SYSTEM TO SUPPORT SOLDIER ASSESSMENTS WITH THE CONSEQUENCES TEST

EXECUTIVE SUMMARY

Research Requirements:

Measures of divergent thinking and creativity, such as the Consequences Test, have been shown to predict important aspects of leader performance in the military. These scales describe unique situations and require examinees to list implications that might arise from those situations. Expert panels have been required to score these measures and assess the creativity and diversity of an individual's responses. However, the use of expert panels has made these scales impractical for most large-scale testing applications. To address this limitation, Latent Semantic Analysis (LSA) techniques were used to compute scores reflecting the creativity and diversity of responses for the creativity test. Analyses demonstrated that the LSA scores were highly correlated with conventional scores, reaching a correlation of 0.94 with human raters, and were moderately correlated with performance criteria. This approach to scoring measures of divergent thinking and creativity solves many practical problems, including the time for humans to rate open-ended responses and the difficulty in achieving reliable scoring.

To use the Consequences Test in large-scale operational settings for the military, dozens of test items are required to maintain test security and to prevent Soldiers from receiving the same items at multiple testing times over the course of their careers. Having an adequate number of test items requires that additional LSA scoring algorithms be developed for additional Consequences Test items.

Procedure:

We extended our past work developing automated scoring for the test of creativity by creating LSA scoring algorithms for four newly constructed test items. These scoring algorithms were based on the scoring techniques used with the original Consequences Test items, with new techniques incorporated as needed. In addition, we updated the scoring algorithms for the original five test items to allow all of the scoring algorithms to be distributed together for streamlined assessment through an integrated scoring tool.

Findings:

The four new scoring algorithms were validated against subject matter expert (SME) scores used for training the scoring algorithms. The correlations between SMEs and scoring algorithms ranged from 0.76 to 0.91, while the correlations between the two SMEs ranged from 0.89 to 0.99. This result indicates that the scoring algorithms are able to score responses almost as consistently as the SMEs. The updated scoring algorithms for the original five test items show slight improvement in comparison with SME scores compared to the original scoring algorithms.

Utilization and Dissemination of Findings:

The LSA scoring algorithms were provided to the Army as three scoring tools designed to meet different emerging Army needs. Each version included both the scoring algorithms for the original five test items and for the four new Army test items. The initial scoring tool was delivered just in time for the Battalion Commander Assessment Program (BCAP), which was executed in January 2020. Two subsequent scoring tools were developed that 1) improved the user experience and data screening capabilities on a personal computer, and 2) provided a server version of the scoring tool that can be integrated with other assessments into an integrated assessment and scoring solution being developed by the Army.

DEVELOPING AN AUTOMATED SCORING SYSTEM TO SUPPORT SOLDIER ASSESSMENTS WITH THE CONSEQUENCES TEST

CONTENTS

Developing an Automated Scoring System to Support Soldier Assessments with the Consequences Test

CHAPTER 1: INTRODUCTION

Leaders in today's Army, including both commissioned officers and non-commissioned officers, face a wide array of challenges. Current conflicts require leaders to adapt quickly to evolving threats, to regularly engage in creative problem solving, and to assume greater responsibility and independence at more junior levels. Successful leadership must combine versatile decision-making and critical thinking skills with creativity. In order to identify officers and officer candidates who are likely to perform well under these circumstances, an updated approach to personnel selection is required.

Measures of creativity, and particularly divergent thinking, have been shown to be good predictors of important aspects of career performance in the Army, including continuance and progression (Mumford, Marks, Connelly, Zaccaro, & Johnson, 1998; Zaccaro et. al., 2012; Zaccaro et. al., 2015). Evidence shows that creative leaders are crucial for innovation and organization success (Mumford & Hunter, 2005; Mumford, Medeiros, Steele, Watts, & Gibson, 2014). Effective leadership in the Army includes performing major duties with creative aspects. For example, leaders need to create and disseminate a vision of the future, and they need to create an environment that fosters innovative and critical thinking (Paullin, Legree, Sinclair, Moriarty, Campbell, & Kilcullen, 2014).

Perhaps the most commonly used measure of creativity is Guilford's Consequences Test (Christensen, Merrifield, & Guilford, 1953). It contains five items that use an open-ended response format. Each test item describes an unusual situation, and participants are allowed two minutes to list as many consequences to the situation as possible. Scoring the Consequences Test has traditionally required the subjective judgment of Subject Matter Experts (SMEs; Guilford & Guilford, 1980). Human raters score the test by first determining if each response is unacceptable or acceptable. Unacceptable responses include those responses that either duplicate earlier statements or are judged irrelevant to the described situation. Acceptable responses are then categorized as obvious or remote. Obvious responses include immediate consequences of the situation as well as vaguely described implications. Remote responses include consequences that are geographically or temporally distant or that involve a new system. Scores are computed so that superior performance is reflected by the ability to provide a large number of acceptable responses, with remote responses being double-weighted over obvious responses.

Due to the subjective nature of the scoring process, the Consequences Test scoring procedure (Guilford & Guilford, 1980) requires that several SMEs independently assess each protocol, then meet as a team to resolve differences in opinion. In addition, responses must be scored in batches to maintain consistent scoring, rather than being scored individually or in real time. Despite having the potential to predict important aspects of career performance, the complex scoring process is both time-consuming and labor-intensive, making it impractical to administer on a large scale.

Prior Work (2013-2018)

As an alternative to SME ratings, we developed automated scoring for the Consequences Test using Latent Semantic Analysis (LSA). LSA is a machine learning technology that provides the ability to extract and infer the meaning of words from large collections of text (Martin & Berry, 2010; Landauer, Laham, & Foltz, 2003). 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 the example responses: "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 more and skip both imply an absence; more generally, these phrases may have similar meaning within a common paragraph. Based on the analysis of a large corpus of text, 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, LSA would judge these two phrases to be semantically similar.

Past analyses have shown that LSA is useful for assessing the quality of essays that have been written by respondents on specific topics (Landauer, Laham, & Foltz, 2003). LSA technologies have been adapted to support automated essay scoring in educational settings, both to provide writing instruction (Streeter, Berstein, Foltz & DeLand, 2011), and for low and high stakes writing assessments including the SAT, GRE, and GMAT (Shermis, 2014; Zhang, 2013). Analyses also demonstrate that LSA-generated scores often have high agreement with SMEs, comparable to the agreement between SMEs (Landauer, et al., 2003; Shermis, et al., 2010; Shermis, 2014). LSA has also 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 focused on constrained test items that assess content knowledge (Liu, Brew, Blackmore, Gerard, Madhok, & Linn, 2014). Streeter and her colleagues (2011) report that over five years of scoring short answers, up to 50% of the short answer items from a state science test could not be scored accurately enough for high stakes testing. Thus, successfully scoring unconstrained short answer responses requires going beyond current automated scoring capabilities and developing new approaches.

Our past work demonstrated that LSA-based automated scoring was able to approximate SME scoring of Consequences Test items (LaVoie, Parker, Legree, Ardison, & Kilcullen, 2019; LaVoie, Parker, Legree, Ardison, & Kilcullen, 2017). The automated scoring algorithms came very close to SMEs' level of agreement, reaching a correlation of 0.94 with the human ratings. The LSA scores achieved excellent convergence with SME ratings, indicating that an automated scoring algorithm may be used to effectively score a measure of creativity and divergent thinking in lieu of the cumbersome human scoring process. The automated scores also showed very similar patterns of correlations with several personality measures collected from a sample of 1,863 ROTC cadets who participated in the Leadership Development and Assessment Course during the summer of 2013 (LaVoie, Parker, Legree, Ardison, & Kilcullen, 2019; LaVoie, Parker, Legree, Ardison, & Kilcullen, 2017). An automated scoring tool was built to allow new

responses to the Consequences Test to be automatically scored using the developed scoring algorithm.

The scoring tool is a computer program that combines a user interface with the scoring algorithms. This allows a user to submit a data file containing new responses to one or more test items. The program then scores the new responses with the automated scoring algorithm(s), and returns the scores in a file.

New items for the Consequences Test have recently been incorporated into an Army form of the test. The Army would like to be able to automatically score these items, which requires that new scoring algorithms be developed for each of the new items. Furthermore, the Army plans to use these as part of an operational testing program. Because the test will be given to multiple populations, and potentially to the same individuals over time as they progress through their careers, it is crucial to have a wide range of test items available. Having multiple forms using different test items will give the Army the ability to avoid administering the same form to a given individual multiple times and help to maintain test security. Having a large number of test items available requires that scoring algorithms be developed for many more test items, and that responses to the new items be scored as efficiently as possible. To this end, we completed two significant tasks in the first year of this project, which is the focus of the current report. These tasks included: (1) developing automated scoring algorithms for new Army Consequences Test items and additional versions of the scoring tool that incorporate the new algorithms, and (2) creating an improved user interface to facilitate operational use of the Consequences Test. Both tasks, which were completed during the period of September 2019 – September 2020, are discussed in the subsequent sections.

CHAPTER 2: THE EXPANSION OF AUTOMATED SCORING

Subject Matter Expert Ratings

The first step in developing high quality automated scoring algorithms is producing accurate and consistent human scores of the responses to the test items. Subject matter experts (SMEs) scored the responses. A scoring rubric and examples of scored responses helped the SMEs maintain consistency with each other over time. The procedure used to score the new Army items was somewhat different than the procedure used to score the original test items. The most important difference is that the SMEs scoring the new Army test items discussed any discrepancies in order to reach consensus, forcing a high level of agreement among raters. The procedure used to score the original items required the raters to independently score each item prior to discussing the score and reaching consensus. As a result, the high agreement between the pairs of SMEs who scored the new Army items may be artificially inflated. Indeed, the correlations between the pairs of SMEs who rated the new Army items ranged from 0.89 to 0.99, while the correlations between the raters who scored the five original items ranged from 0.72 to 0.83. One potential problem with inflated agreement is that it is not based on patterns in the data, which can make it more difficult to train an accurate scoring algorithm.

Algorithm Development

We developed new scoring algorithms for the four newly developed Army Consequences Test items. The automated scoring system relies on Latent Semantic Analysis (LSA) and requires two components: a background semantic space and a set of scoring algorithms. 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, 2007). A background sematic 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 updated the semantic space that was previously used to score the original test items. The new space is larger and has better coverage of general language, which is important for scoring a test of creativity which elicits a wide range of responses. It includes 192,955 paragraphs of written language and 557,227 words.

The goal of the automated scoring system was to correctly score each participant's aggregate responses. Each response can receive a content score of zero, one, or two (indicating irrelevant, obvious, or remote). If a response duplicates another response or the examples provided in the text, then it is discounted and does not get combined when calculating the aggregate score. In our past work, we developed separate automated scoring processes for each of these scales: content scoring and identification of duplicates.

For the current scoring algorithms, we focused on developing scoring algorithms that could score the combined responses from each participant. This process offers a simplified technique for scoring responses to the test items. We implemented several new measures designed to improve scoring accuracy and efficiency. These measures were combined using

regression models to develop the scoring algorithms. Many of the new measures are based on LSA. For example, neighborhood metrics were calculated by comparing the semantic similarity of the new, to-be-scored, responses with responses that were previously scored by the SMEs. The most semantically similar responses, called neighbors, can be used to estimate the score that SMEs would give to the new responses. Projecting all of these responses into the background space allowed us to make semantic comparisons and identify the most similar scored responses. By examining the SME scores for the similar responses, we can estimate an appropriate score for a new item.

In order to build a single scoring tool that incorporates all of the Consequences Test items, including the four new items and the five original items, we also developed updated scoring algorithms for the five original items. These scoring algorithms applied the new, simplified scoring technique to score the original items.

Results

The nine scoring algorithms were evaluated by comparing them to the SME Consequences Test ratings for each item. Because the scoring algorithms were developed based on these same ratings, a direct comparison can result in an inflated estimate of the generalizability of the scoring algorithms to new data. To avoid this problem, we used three hold-out sets. The hold-out sets were created by randomly selecting one half of the data set and using it for training, leaving the remaining half for testing. For each set, the Pearson correlation was calculated by comparing the automated score with the SME consensus score. These correlations were compared to the correlations between the two SME ratings. All results were calculated by averaging across hold-out sets.

The correlations between the SME consensus ratings and the scoring algorithms scores for the four new Army test items ranged between 0.76 and 0.91, while the correlations between the SMEs ranged from 0.89 to 0.99. The correlations, averaged across hold-out sets, for each item are shown in Table 1.

Table 1
Correlations between Subject Matter Expert scores and Computer scores for the four new Army Consequences Test items.

Army Consequences Test Items	Correlation Between SMEs and ASAs	Correlation Between SMEs
Q3	0.76	0.99
Q4	0.79	0.89
Q6	0.88	0.93
Q7	0.91	0.96

Note. SME = Subject Matter Experts, ASA = Automated Scoring Algorithms. Sample sizes for each item are Q3 n=125, Q4 n=125, Q6 n=121, Q7 n=121.

The correlations between the SME consensus score and the original test items, averaged across hold-out sets, are shown in Table 2 for both the original scoring algorithms and the improved scoring algorithms. The average correlation between the SME scores and the LSA scoring algorithm scores across all five items improved from 0.85 to 0.88 with the new scoring algorithms. Note that the question numbers have some overlap as the two sets of items were developed independently.

Table 2

Correlations between SME scores and Computer scores for the five original Consequences Test items for the original ASAs and the new ASAs.

Original Consequences Test Items	Original Consequences Test Scoring	New Consequences Test Scoring
Item O-1	0.86	0.88
Item O-2	0.84	0.86
Item O-3	0.86	0.88
Item O-4	0.88	0.90
Item O-5	0.80	0.87

Note. SME = Subject Matter Experts, ASA = Automated Scoring Algorithms. Samples sizes for each item are: O-1 n=709, O-2 n = 705, O-3 n=708, O-4 n=699, O-5 n=706.

CHAPTER 3: DEVELOPMENT OF CONSEQUENCES TEST SCORING TOOLS

In order to use the scoring algorithms to automatically score responses to the Consequences Test, the Army requires a scoring tool capable of taking a data file containing new responses, automatically scoring these responses, and producing a data file including the scores. The first version of the scoring tool was delivered at the end of 2019, just in time for the Battalion Commander Assessment Program (BCAP), which was executed in January 2020. This first version included only new scoring algorithms for the four new Consequences Test items, although the original scoring algorithms for the original five test items were included on the same computer.

The second version of the scoring tool improved the user experience and data screening capabilities on a personal computer and included both the scoring algorithms for the four new items and updated scoring algorithms for the original five items. This scoring tool replaced the first version of the scoring tool as the preferred tool for scoring responses by an individual researcher. This tool uses drag-and-drop functionality to score a new data file. The user simply drags a response file with test item responses over the scoring tool icon and the responses are scored. The response file must be formatted correctly. The preferred format is a comma separated file (.csv). The headers in the .csv file must begin with Roster Number and then DOD ID followed by the item number with each response in a separate column. For example, the response column headers might look like this for item five: Q5_1, Q5_2, Q5_3, Q5_4. There is no limit to the number of responses that may be included. The scoring tool can score responses from 800 participants in less than ten minutes. The scores are appended to the response data file, submitted by the user, for easy reference.

The third version of the updated scoring tool provided a server version of the scoring tool that can be integrated with other assessments into an integrated scoring solution being developed by the Army. This version also included both the scoring algorithms for the four new items and for the original five items. This version, to be run on the server, has been delivered but has not yet been installed and tested on an Army server, nor has it been integrated into the end-to-end assessment tool the Army has planned. We anticipate that our work on this version will continue until these tasks have been completed.

Future plans for both scoring tools, the personal computer and server versions, include updates to add more scoring algorithms for items as they are developed under future projects.

CHAPTER 4: DISCUSSION

The updated scoring algorithm development procedure worked well and allowed us to develop new scoring algorithms and update the original ones. This in turn meant we were able to develop an integrated scoring tool that included all of the scoring algorithms developed to date. The new scoring algorithms achieved high correlations with the SME ratings, approaching the agreement between SMEs and providing an appropriate level of accuracy for an operational test. The Army now has several scoring tools available to allow the quick and effective scoring of new responses to the Consequences Test items.

In future work, we plan to address one of the key limitations of automated scoring – the psychometric properties of the human scoring used to train the scoring algorithms. Using a process we term "human scoring calibration", trained human raters will score a sample of responses, and then we will evaluate the concurrent and predictive validity of the human scores. Following this, the rubric and anchors will be updated. The process will be repeated: the human raters will score a new sample of responses and concurrent and predictive validity will be evaluated, until the concurrent and predictive validity have reached desired levels. In this way the rubrics, anchors, and human scoring will be calibrated prior to scoring all of the responses. The scoring algorithms will then be trained on the human scores. Because scoring algorithms are generally limited by the qualities of the human scoring, this should allow us to develop scoring algorithms with higher concurrent and predictive validity.

Another consideration for improving the human scoring of responses is whether the process of independent ratings, followed by discussion to reach consensus, is necessary. A more efficient process would be to simply average the scores from the two independent raters. There is also reason to be concerned that some pairs of raters may not have scored fully independently as the correlations between raters were very high. Because this can make scoring algorithms more difficult to train and can reduce their overall agreement with the SME ratings, we plan to use a computer scoring tool in the future to help ensure that raters make their judgments independently.

In order to improve the operational implementation of the Consequences Test, we anticipate developing many more scoring algorithms to allow for parallel forms of the test to be used across a Soldier's career. Test security is of particular importance as this test is now being used in the Battalion Commander and Company Commander Assessment Programs, making it likely that the same Soldiers could encounter these specific items multiple times, possibly inflating their scores, if there are not sufficient items available. These additional scoring algorithms will also need to be included in future scoring tools.

Future plans for this project include evaluating the validity of automated scores as a measure of creativity in leaders. This is particularly important as the test is being used operationally. As appropriate data become available from future data collections, we plan to examine correlations among creativity scores by rank using samples from commissioned and non-commissioned officers. We hope to link scores with cadet/Officer performance metrics from future data collections as was done for the original Consequences Test items (LaVoie, Parker,

Legree, Ardison, & Kilcullen, 2019). We will also examine the validity by gender, race, and ethnicity to ensure that no systematic bias exists in the automated scoring of this test.

REFERENCES

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

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

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

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*(3), 295-308.

LaVoie, N., Parker, J., Legree, P., Ardison, S., & Kilcullen, B. (2019). Using Latent Semantic Analysis to score short answer constructed responses: Automated scoring of the Consequences Test. *Educational and Psychological Measurement, 80*(2), 399-414.

LaVoie, N., Parker, J., Legree, P., Ardison, S., & Kilcullen, B. (2017). Automated scoring of the Consequences Test using Latent Semantic Analysis. Poster presented at the SIOP conference, Orlando, FL.

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.

Mumford, M. D. & Hunter, S. T. (2005). Innovation in organizations: A multi-level perspective on creativity. *Research in Multi-Level Issues, 4,* 11-73.

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*(2), 151-163.

Mumford, M. D., Medeiros, K. E., Steele, L., Watts, L. L., & Gibson, C. (2014). Leadership, creativity, and innovation: An overview. In M. D. Mumford (Ed.), *Leadership, creativity, and innovation*. Thousand Oaks, CA: Sage.

Paullin, C., Legree, P. J., Sinclair, A. L., Moriarty, K. O., Campbell, R. C., & Kilcullen, R. N. (2014). Delineating officer performance and its determinants. *Military Psychology, 26*(4), 259-277.

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.

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., Gilrane, V.L., Robbins, J.M., Bartholomew, L.N., Young, M.C., Kilcullen, R.N., Connelly, S., & Young, W. (2012). *Officer individual differences: predicting long-term continuance and performance in the U.S. Army* (ARI Technical Report 1324). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Zaccaro, S. J., Connelly, M. S., Repchick, K. M., Daza, A. I., Young, M. C., & Kilcullen, R. N. (2015). The influence of higher order cognitive capacities on leader organizational continuance and retention: The mediating role of developmental experiences. *The Leadership Quarterly, 26,* 342–358.

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