Research Note 2022-02

Development of an Automated Scoring System to Support Soldier Assessment with the Consequences Test

Noelle LaVoie
James T. Parker
Amy Santamaria
Parallel Consulting

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

December 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 Reviews by:

Melissa J. Glorioso, U.S. Army Research Institute
Karly M. Schleicher, U.S. Army Research Institute

DISPOSITION

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
December 2021	Final	August 2019 – September 2021

4. TITLE AND SUBTITLE

Development of an Automated Scoring System to Support Soldier Assessment with the Consequences Test

5a. CONTRACT NUMBER
W911NF-19-C-0097

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
622785

6. AUTHOR(S)
LaVoie, Noelle
Parker, James
Santamaria, Amy
Young, Mark C.
Legree, Peter J.

5d. PROJECT NUMBER
A790

5e. TASK NUMBER

5f. WORK UNIT NUMBER
1011

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Parallel Consulting, LLC
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 & Social Sciences
6000 6TH Street (Bldg. 1464 / Mail Stop 5586)
Fort Belvoir, VA 22060-5610

10. SPONSOR/MONITOR'S ACRONYM(S)
ARI

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Research Note 2022-02

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

13. SUPPLEMENTARY NOTES
ARI Research POC: Dr. Mark C. Young, Selection and Assignment Research Unit

14. ABSTRACT

Measures of creativity, and particularly divergent thinking, such as the Consequences Test (Guilford & Guilford, 1980), have been shown to be good predictors of important aspects of career performance in the Army, including continuance and progression. These scales describe unique situations and require examinees to list implications that might arise from those situations. 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. In the latest two-year effort, we developed new scoring models for seven additional test items, increasing the available number of test items in order to improve test security. We also developed new scoring tools that incorporates the five original test items, whose scoring models were developed in an earlier project in 2013-2018, with the new seven items. The scoring models perform well, and the new scoring tool provides an alternative to human scoring that can be integrated with other assessment programs.

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

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Tonia S. Heffner
Unclassified	Unclassified	Unclassified	Unlimited Unclassified	22	19b. TELEPHONE NUMBER 703 545 4408

Research Note 2022-02

Development of an Automated Scoring System to Support Soldier Assessment with the Consequences Test

Noelle LaVoie
James T. Parker
Amy Santamaria
Parallel Consulting

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

Selection and Assignment Research Unit
Tonia S. Heffner, Chief

December 2021

ACKNOWLEDGEMENTS

Additional thanks to the following individuals who contributed to this project: CSM (Ret.) Melissa McFrazier and SGM (Ret.) Derek Johnson.

DEVELOPMENT OF AN AUTOMATED SCORING SYSTEM TO SUPPORT SOLDIER ASSESSMENT WITH THE CONSEQUENCES TEST

EXECUTIVE SUMMARY

Research Requirement:

Measures of divergent thinking, such as the Consequences Test (Guilford & Guilford, 1980), 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 to 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 Consequences Test. Analyses demonstrated that the LSA scores were highly correlated with conventional scores. 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.

The current effort was designed to address some of the practical requirements for implementing the Consequences Test in large-scale operational settings in the military. In this context, dozens of test items are needed 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:

During this two-year effort, we extended our past work developing automated scoring for the test of creativity by building LSA scoring algorithms for seven 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 that had been developed under a previous effort in 2013-2018. We integrated all of the scoring algorithms, five original and seven new, into a scoring tool for streamlined assessment.

Findings:

The seven 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 as a series of scoring tools designed to meet different emerging needs. Each version included both the scoring algorithms for the original five test items and for the new Army test items. The initial scoring tool was delivered just in time for the Battalion Command Assessment Program (BCAP), which was executed in January 2020. This tool included the five original items along with four new ones. Two subsequent scoring tools were developed that 1) improved the user experience, and 2) provided a server version of the scoring tool that can be combined with other assessments into an integrated assessment and scoring solution being developed by the Army. At the end of Year 2, updated versions of both the personal computer and the server scoring tools were delivered that included the automated scoring algorithms (ASAs) for the original five items, the four new items from Year 1, and three additional new items developed in the final project year.

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

CONTENTS

List of Tables

DEVELOPMENT OF AN AUTOMATED SCORING SYSTEM TO SUPPORT SOLDIER ASSESSMENT WITH THE CONSEQUENCES TEST

Background

Leaders in today's Army, including both commissioned 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 divergent thinking, an aspect 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, such as a new form of government or type of regulation. 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 five of the original items that are part of 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. In contrast, "key word search" algorithms look for exact matches to specific words. 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, Bernstein, 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 Burstein, Higgins, & Zechner, 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. This is despite the fact that efforts are made to constrain the potential correct responses to as small a set as possible. Responses to the Consequences Test are much less constrained, as it is a measure of divergent thinking and intentionally elicits as wide a range of responses as possible. Thus, successfully scoring unconstrained short answer responses goes beyond current automated scoring capabilities and requires developing new approaches. In particular, we applied techniques that are more commonly applied to automatically scoring essays, such as LSA-based measures of neighborhood similarity and coherence.

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 divergent thinking in lieu of the cumbersome human scoring process. The automated scores also showed very similar patterns of correlations with several measures collected from a sample of 1,863 Reserve Officers' Training Corps (ROTC) cadets who participated in the Leadership Development and Assessment Course during the summer of 2013 (LaVoie et al., 2019; LaVoie et al., 2017). These measures were from the Cadet Background and Experiences Form, a multiple-choice questionnaire assessing past behaviors and experiences that are related to officer performance and retention (Putka, 2009) and included scales such as Achievement Orientation, Army Identification, Fitness Motivation, Peer Leadership, and Stress Tolerance. Leader Knowledge Test Characteristics and Skills were also measured and two outcome measures were included: Cadet Grade Point Average, and Cadet Order of Merit Listing.

Extending the Consequences Test with ARI Items

New items for the Consequences Test have recently been developed by ARI. These items are in addition to the items developed for the original Consequences 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 along with the associated algorithms.

In the first year of this project, we developed automated scoring algorithms for four new Army Consequences Test items (referred to here as Q3, Q4, Q6, and Q7), along with additional versions of the scoring tool that incorporates these new algorithms. We also created an improved user interface to facilitate operational use of the Consequences Test. In the second year of this project, we supervised expert human scoring for three more new Army Consequences Test items (referred to here as Q9, Q10, and Q11), we developed automated scoring algorithms for them, and we incorporated these algorithms into the personal computer and server versions of the scoring tool. Note that the items numbers are arbitrary and are used to be consistent with internal ARI documents as well as to maintain the convention used in the delivered scoring tools.

Method

Participants

Consequences Test responses were collected from Army participants across multiple research projects, with participants ranging in rank from officer candidates to full colonel (pay grade O-6). Responses for Questions 3 and 4 (Q3 and Q4) were collected from a total of 531 participants. These participants included O3 and O4 officers pursuing job reassignment and officers enrolled in Army intermediate level education (ILE) and senior service college (SSC) courses. Responses for Q6 and Q7 were collected from a total of 1,026 participants. These participants included O3 and O4 officers pursuing job reassignment, officers enrolled in Army intermediate level education and senior service college (SSC) courses, and ROTC cadets. Responses for Questions 9-11 were collected from between 415 (Q10 & Q11) and 759 (Q9) participants attending senior

service college courses, and ROTC cadets. All data was anonymized, so participant demographic information is not available.

Materials

The Consequences Test consists of 5 items designed to elicit responses requiring creative or divergent thinking and allows the respondent to provide up to 20 responses in 2 minutes. An example item and possible responses are:

What would be the result if people no longer needed to eat?

1. No more dinner
2. Skip breakfast
3. No more cooking shows

The test was administered via the internet as part of larger data collections and a computer program controlled the timing so that each participant was required to consider each item for 2 minutes without ability to move to the next item.

Although 12 items have been developed for the Consequences Test, no participant received more than 5 items. Consequences Test items for Year 1 and Year 2 are referred to as Q3, Q4, Q6, Q7, Q9, Q10, and Q11. In developing our scoring tool, we worked with these seven items and also included the five original Consequences Test items (Guilford & Guilford, 1980), referred to as O-1, O-2, O-3, O-4, O-5 (O = original).

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 subject matter experts (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 Process. Developing high quality automated scoring algorithms begins with producing accurate and consistent human scores to the test item responses. In Year 1, ARI oversaw the human scoring process, and in Year 2, Parallel Consulting oversaw the human scoring process. In all cases Army subject matter experts (SMEs) scored the responses.

Scoring the Consequences Test. Each response is scored on content, and can receive a 0 (for irrelevant responses), 1 (for obvious responses), or 2 (for remote responses), and the participant's score on the item is the sum of the individual response scores. For human raters, each response is also scored as either a duplicate (essentially repeating another response) or not a duplicate. If it is scored as a duplicate, the response is zeroed out and does not count towards the participant's total score.

Considering the example above, the first answer gets 1 point for an obvious response. The second answer gets no points because it is a duplicate of the first answer. The third answer gets 2

points for a remote response. Therefore, this participant's total score for this question is 3 (1 + 0 + 2).

We created a detailed scoring rubric based on Guildford & Guildford (1980) and a list of anchors, or sample responses, for each score (irrelevant, obvious, and remote). The rubric and anchors helped the SMEs maintain consistency with each other over time.

Human Scoring Procedure: Year 1. In Year 1, ARI subject matter experts (SMEs) scored the test item responses. As previously noted, developing high-quality automated scoring begins with producing accurate and consistent human scores. 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.

Human Scoring Procedure: Year 2. In Year 2, a pair of SMEs independently scored Consequences Test responses prior to discussing the score and reaching consensus. This was consistent with the way that the five original items from prior work (2013-2018) were scored. To streamline the human scoring process, we built a web-based scoring tool, called Rate It. SMEs scored all responses to each item before moving on to another item. First, the two raters worked individually on a batch of responses. The tool showed each rater the responses for one participant at a time with two columns to fill in. The first column, content, could be scored 0 (irrelevant), 1 (obvious), or 2 (remote). The second column, duplicate, could be scored as either 1 (unique) or 0 (duplicate). After individually scoring responses for a set of participants, the two raters came together to discuss and reach a complete consensus rating for each response. The Rate It tool showed each rater their own scores and the other rater's scores, highlighting any that differed in either the content or duplicate columns. For scores that differed, the raters discussed their reasoning for the score they gave and arrived at a consensus score, which they input into the Rate It tool. The tool recorded both the individual ratings and the consensus ratings.

SMEs started out scoring each item with very small batches of participants (1 to 5), working while researchers supervised and were available to answer questions. Then SMEs moved on to finish a set of 20 on their own. They continued to work in sets of 20 participant responses until they reached an acceptable level of agreement and felt comfortable moving to sets of 50 participant responses. For the first item, Q9, the raters needed two practice sets of 20 participant responses to reach an acceptable level of agreement and they completed an additional set of 20 before moving to sets of 50. For the next two items, Q10 and Q11, the raters only needed one set of 20 participant responses and then moved on to sets of 50. We checked agreement for the raters

after they completed consensus for each set. We also organized the anchor list into categories and added new anchors, according to feedback from raters. The procedures we followed in Year 2 allowed us to improve the efficiency of the human scoring process while maintaining effective levels of agreement between raters.

Automated Scoring Process. The goal of the automated scoring system was to correctly score each participant's aggregate responses, or total score, for each item. Each response can receive a content score of zero, one, or two (indicating irrelevant, obvious, or remote). We developed new scoring algorithms for the seven newly developed Army Consequences Test items (four in Year 1 and three in Year 2). 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, McNamara, Dennis, & Kintsch, 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 divergent thinking which elicits a wide range of responses. It includes 192,955 paragraphs of written language and 557,227 words.

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 aggregate responses from each participant. This process offers a more efficient 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, these included neighborhood metrics 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 these 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 seven new items and the five original items, we also developed updated scoring algorithms for the five original items. These scoring algorithms applied the new, more efficient scoring technique to score the original items.

Results

For both Year 1 and Year 2, we evaluated the scoring algorithms by comparing 1) how well scores from the SME raters agreed with each other and 2) how well the scoring algorithms agreed with the raters' scores. Agreement between the raters was calculated for the total Consequences Test score (by participant) using Pearson correlations for the SME ratings prior to

their reaching consensus. The 12 scoring algorithms (five original Consequences Test items, four Year 1 new items, and three Year 2 new items) were evaluated by comparing them to the SME Consequences Test ratings for each item. Because the scoring algorithms were trained on these same ratings a direct comparison can result in an inflated estimate of the scoring algorithm performance. To avoid this problem, and better estimate the generalizability of the scoring algorithms to new responses, we used three hold-out sets per item. 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.

For Year 1, the correlations between the SME consensus ratings and the automated scoring algorithm scores for the four new Army test items (Q3, Q4, Q6, and Q7) ranged between 0.76 and 0.91, while the correlations between the SMEs ranged from 0.89 to 0.99. Note that because the aggregate, total score for each item is compared, the agreement is higher than it is at the individual response level. The correlations, averaged across hold-out sets, for each item are shown in Table 1. This performance is consistent with our prior work, and that of other researchers, and is discussed further in the discussion section.

Table 1. Correlations between Subject Matter Expert scores and Computer scores for Year 1 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.
Hold out samples sizes for each item are Q3 n=125, Q4 n=125, Q6 n=121, Q7 n=121.

For Year 2, the correlations between the SME consensus ratings and the automated scoring algorithm scores for the three new Army test items (Q9, Q10, and Q11) ranged between 0.82 and 0.86. All of these models excluded count variables. The correlations between the SMEs ranged from 0.97 to 0.98. The correlations, averaged across hold-out sets, for each item are shown in Table 2.

Table 2. Correlations between Subject Matter Expert scores and Computer scores for Year 2 Army Consequences Test items.

Army Consequences Test Items	Correlation Between SMEs and ASAs	Correlation Between SMEs
Q9	0.86	0.98
Q10	0.82	0.97
Q11	0.84	0.98

Note. SME = Subject Matter Experts, ASA = Automated Scoring Algorithms.
Hold out samples sizes for each item are Q9 n=390, Q10 n=251, Q11 n=246.

The correlations between the SME consensus score and the original test items, averaged across hold-out sets, are shown in Table 3 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 3. Correlations between Subject Matter Expert 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.
Hold out 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

Tool Development

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. In Year 1, we built three versions of a scoring tool.

The first version of the scoring tool was delivered at the end of 2019, just in time for the Battalion Command Assessment Program (BCAP), which was executed in January 2020. This first version included only new scoring algorithms for the four new Consequences Test items

(Q3, Q4, Q6, and Q7), although the scoring tool that was provided to score the original five test items was installed 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 and includes sample data files that provide examples of the appropriate formatting for an input data file. Once the data file is prepared, 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. As mentioned previously, the scoring tool is more efficient than the one delivered in the previous project for the original items. This 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 with clear labels for the new scores.

The third version of the updated scoring tool was a Unix server version of the scoring tool that can be combined 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 was delivered at the end of Year 1, ready to be installed and tested on an Army server.

In Year 2, we updated the second and third versions of the scoring tool to add scoring algorithms for three new Consequences Test items (Q9, Q10, and Q11). These versions were delivered at the end of Year 2. The personal computer version was delivered on laptops while the server version was delivered on DVD, ready to be installed and tested on an Army server.

In all cases, ARI personnel received training on how to use the scoring tool including preparing input data files, scoring the data, and using the output files with the automated scores.

Discussion

In this project, we have built on previous work to develop automated scoring algorithms using Latent Semantic Analysis. This has allowed us to efficiently score the Consequences Test, providing an alternative to the time-consuming and impractical approach of using human scoring at a large scale. In Year 1, we developed scoring algorithms for four new Consequences Test items and updated scoring algorithms for the original five items. We used this to develop an integrated scoring tool that included scoring algorithms for those nine items. 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 work in Year 1 provided the Army with several scoring tools to allow the quick and effective scoring of new responses to the Consequences Test items. In Year 2, we supervised human scoring of three

more new Consequences Test items and developed scoring algorithms for those items. We then incorporated these three new items into the scoring tool, delivered as an updated server version of the tool. The Army now has a scoring tool available to allow the quick and effective scoring of 12 Consequences Test items.

The scoring algorithms achieved good convergence with SME ratings, supporting previous work that demonstrated that an automated scoring model may be used to effectively score a short answer measure of divergent thinking in lieu of the cumbersome human scoring process.

Limitations and Future Work

One notable limitation of the current scoring tools is that they are based on the limited sample of data provided for training. Incidents of cheating or other possible attempts to game the system are not present in this data, and duplicate responses occur in only 6.4% of the responses. As a result, the scoring algorithms have not been trained to recognize and exclude responses that contain unexpected content (e.g., nonsense syllables, random strings of letters, emojis or foreign languages). The automated scoring algorithms are trained to score aggregate responses, rather than individual responses from each participant. This means that the scoring tool does not explicitly exclude duplicate responses. Rather, it matches the response it is scoring with previously scored responses to estimate the appropriate score. This value is then combined with other measures using the regression procedure to predict the score. We recommend adding filters to the scoring system that can detect a variety of possible issues with the data. The filters would be algorithms that the data is run through to screen the data for potential issues prior to scoring. For example, a filter to identify duplicate responses could be added to ensure that if the tool encounters responses that include a higher-than-expected number of duplicates the tool will exclude these and lower the score appropriately. Another filter could be added to identify unusual responses that are dissimilar to the responses on which the algorithms were originally trained. These could be flagged for human review in order to identify new types of cheating or inappropriate responses that are being submitted to the Consequences Test. The incidence of unusual responses could also be monitored to determine if the responses have drifted in content, length, or other dimensions from the original responses on which the algorithms were trained. If, for example, responses to a particular item begin to look dissimilar to the original training responses it may be advisable to build a new scoring algorithm to maintain high scoring accuracy. Note that it is a limitation of machine learning that the systems are unable to provide accurate scores on data that is too dissimilar to the data included in training.

Plans for future work should include further validation of the automated scoring of the Consequences Test. When the Consequences Test is used operationally, it is important to evaluate the validity of automated scores as a measure of divergent thinking in leaders. Future research should examine correlations among divergent thinking scores by rank using samples from commissioned and non-commissioned officers. It would be appropriate to link scores with

cadet/officer performance metrics from future data collections, as was done for the original Consequences Test items (LaVoie et al., 2019). We also recommend examining validity by gender, race, and ethnicity to ensure that no systematic bias exists in the automated scoring of this test.

In order to improve the operational implementation of the Consequences Test, we recommend 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 Command and Colonels Command 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 work should also examine the relationships among these new ARI items and their associated scoring algorithms to determine which items are best combined into a variety of parallel forms of the test. For example, intercorrelations among items, item reliability, and the minimum number of items needed to maintain reliability with both SME and automated scoring should all be evaluated.

Conclusion

Automated scoring is a viable alternative to time consuming hand scoring of the Consequences Test. This approach to scoring the Consequences Test 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. As a result, it is now more practical to use the Consequences Test in large scale data collections for purposes including personnel selection.

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. (2017). Automated scoring of the Consequences Test using Latent Semantic Analysis. Poster presented at the SIOP conference, Orlando, FL.

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.

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* (3195-3204). New York: Taylor & Francis.

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

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20* (1), 53-76.

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_