Building Automated Assessments of Interpersonal Leadership Skills

Randy Brou, PhD
Army Research Institute, Fort Benning
Fort Benning, GA
randy.j.brou.civ@mail.mil

Gary Stallings, Sean Normand
Northrop Grumman Systems Corporation
Columbus, GA
gary.stallings@ngc.com, sean.normand@ngc.com

Blake Ledford
Consortium Research Fellows Program
Fort Benning, GA
cameron.b.ledford.ctr@mail.mil

Ian Stearns
Northrop Grumman Systems Corporation
Albuquerque, NM
ian.stearns@ngc.com

ABSTRACT

Producing effective leaders is a concern for training departments across the military, industry, and academia. The specific skill requirements for leaders across these domains varies, but effectively interacting with people is a requirement in any leadership role. Despite the broad utility of interpersonal leadership skills, methods available to systematically assess those skills are limited. Some organizations rely on self-report measures or situational judgment tests of leadership skills. Others may use performance measures gathered during observations of live assessments. The former set of methods is disadvantaged by social desirability bias and ability to identify criteria distorting participant responses. The latter set of methods is costly in time and human resources, and may suffer from observer subjectivity. The current research investigated another option for assessing interpersonal leadership skills: reactive, computer-based scenarios using unprompted, natural language responses as inputs. This method helps to mitigate the problems of self-report measures and may be widely used at a fraction of the costs associated with live assessments, but it faces two challenges. First, the assessment tool must be able to interpret natural language responses accurately. Second, virtual agent behaviors must be flexible enough to believably react to unguided inputs. In an experiment, US Army Officer Candidates interacted with virtual agents representing leaders, peers, and subordinates in three scenarios composed of 4 to 7 related vignettes. Free-text responses provided during real-time conversations with the agents influenced the outcomes of each scenario. Interactions with the agents were analyzed to determine if the assessment method could accurately detect differences in interpersonal leadership skills among Officer Candidates. Results of this research provided initial evidence that such differences can be detected using the experimental method. Further, results provided insights into the amount of training data needed for language libraries to accurately interpret unprompted inputs and for developing sufficiently flexible agents.

ABOUT THE AUTHORS

Randy Brou, PhD is a Research Psychologist in the Army Research Institute, Fort Benning Research Unit. He has 19 years of experience in conducting applied research for the Department of Defense. His work has focused on training effectiveness and the measurement of individual and team attributes relevant for successful performance.

Gary Stallings is a Behavioral Research Scientist at Northrop Grumman Systems Corporation, where he provides research and analysis for behavioral science research projects in support of the U.S. Army.

Sean Normand is a Program Manager at Northrop Grumman Systems Corporation, where he manages and leads the technical planning and execution of multiple behavioral science research projects in support of the U.S. Army.

Ian Stearns is a Software Engineer at Northrop Grumman Systems Corporation. He has worked on multiple web, database, and business applications, and now helps with various analytics projects.

Blake Ledford is a Research Assistant at the Army Research Institute, Fort Benning Research Unit, where he assists with research design and software development.

Building Automated Assessments of Interpersonal Leadership Skills

Randy Brou, PhD
Army Research Institute, Fort Benning
Fort Benning, GA
randy.j.brou.civ@mail.mil

Gary Stallings, Sean Normand
Northrop Grumman Systems Corporation
Columbus, GA
gary.stallings@ngc.com, sean.normand@ngc.com

Blake Ledford
Consortium Research Fellows Program
Fort Benning, GA
cameron.b.ledford.ctr@mail.mil

Ian Stearns
Northrop Grumman Systems Corporation
Albuquerque, NM
ian.stearns@ngc.com

INTRODUCTION

Organizations depend on skilled leadership, but becoming an effective leader depends on mastering many skills (Mumford, Campion, & Morgeson, 2007; Mumford, Zaccaro, Harding, Jacobs, & Fleishman, 2000). Mumford et al. (2007) delineate four domains of leadership skills: cognitive, business, strategic, and interpersonal. Ideally, organizations would have clearly defined methods and metrics for assessing skill progression across each of these domains as part of their approach for training prospective leaders. Such methods and metrics are readily available for technical skills training (e.g., tactics, market analysis) which correspond to the first three skill domains; however, systematically assessing interpersonal skills has proven to be an obstacle (Bedwell, Fiore, & Salas, 2014). The current research focused on the development of a novel methodology for the systematic assessment of interpersonal leadership skills in the context of US Army Officer training.

Interpersonal Leadership Skills

Army leaders face the unique challenge of inspiring their subordinates to persevere in life-threatening conditions, often on foreign soil and during prolonged periods of stress. Additionally, leaders must instill a specific set of ethics and values in their subordinates which serve to guide their conduct and decision making. The behaviors associated with effectively meeting these demands constitute interpersonal leadership skills. The US Army Officer Candidate School (OCS) is one of the organizations charged with the training and development of future Army leaders. OCS uses the Army Leadership Requirements Model (FM 6-22) as a guiding framework for understanding the interpersonal skills of interest for Army Officers. The current research focused on developing a framework for systematically assessing a subset of the skills defined in FM 6-22. Table 1 provides abbreviated definitions of the interpersonal leadership skills targeted by the current research.

Table 1. Targeted Interpersonal Leader Skills

Skill	Definition
Leads Others	Motivates, inspires and influences others to take initiative, work towards a common goal, and achieve objectives by setting an example for others, serving as a role model, and maintaining high standards in all aspects of behavior and character
Develops Others	Prepares others for success by encouraging and supporting them to grow as individuals and teams
Creates a Positive Environment	Creates a positive cultural and ethical environment by establishing conditions of effective influence
Communicates	Clearly expresses ideas to ensure understanding, actively listens to others, and practices effective communication techniques
Gets Results	Produces consistent results by developing and executing plans that provide team members with clear direction, guidance, and priorities towards mission accomplishment

The skills associated with effective interpersonal leadership are intangible, making them inherently difficult to capture using the same traditional assessment methods that are effective for technical skills assessment. Historically, the Army has relied on two methods for interpersonal leadership assessment: self-report and live assessment. Unfortunately, both of these methods have short-comings limiting their effectiveness in assessing interpersonal skill proficiency. To address these limitations, the current research investigates a third option for assessing interpersonal skills: reactive, open-response assessments.

Self-Report Assessments

Early uses of self-report measures in the Army date back to World War II (e.g., Stouffer, Suchman, DeVinney, Star, & Williams, 1949). Self-report inventories have been used to assess interpersonal leadership skills and other "soft" skills over the years. Self-report measures benefit from being easy to administer and are relatively inexpensive, but have several weaknesses. For example, social desirability bias may impact self-report responses via respondents' need for social acceptance or backgrounds (Donaldson & Grant-Vallone, 2002; Nederhof, 1985). Further, Niessen, Meijer, and Tendeiro (2017) noted that "self-presentation" attenuates the predictive validity of self-report assessments administered within high-stakes contexts (e.g., for job selection). Respondents' ability to identify criteria (ATIC) may also influence the way they respond to self-report assessments (Klehe, et al., 2012). For instance, Krumm, et al. (2015), showed that respondents were able to select the "best" answers to situational judgment tests (another common assessment methodology falling into the self-report category) about 70% of the time, even when they were only supplied with the answer choices and not the actual situation stems. That is, respondents were able to examine answer choices, identify the skill or attribute (criteria) under examination, decide which choice most effectively demonstrated that criteria, and make the "best" selection regardless of the degree to which it accurately reflected their own skills or attributes. Given these issues with self-report assessments, the Army tends to rely more heavily on live assessments to identify interpersonal skill gaps.

Live Assessments

Field training exercises (FTX) are a common training and assessment methodology employed by the Army. FTX are typically limited scope, scenario-driven events that allow instructors to watch trainees practice specific skills in a realistic setting. These assessments tend to be more objective than self-report measures because trainees are executing the skills; however, some subjectivity on the part of the instructors/observers may exist at times. With respect to interpersonal skills, live assessments offer Army instructors the opportunity to see trainees interacting with other Soldiers while under stress, and/or while in situations that have been designed to challenge interpersonal leadership skills. The U.S. Army Combined Arms Center, for example, describes an FTX wherein an incoming commander was challenged with understanding the dynamics of an area of operation he was inheriting and developing/executing a plan for dealing with enemy forces in the area. The commander initially displayed enthusiasm for the mission and encouraged his subordinates, but was dismissive of the inputs he received from some of his Soldiers during mission planning and did not foster a positive environment in the unit. This deficiency in interpersonal leadership skills led to negative consequences in the unit's performance. Although live assessments provide valuable opportunities to observe skill gaps, the unfortunate reality is that due to the number of Soldiers requiring training at any given time and resources required to conduct a live training exercise, most Army courses can only reasonably evaluate any given Soldier once or twice in such a setting. Given the complexity of the skills in question, having more opportunities to objectively assess skill progression would be beneficial.

Reactive, Open-Response Assessments

An ideal assessment methodology would combine the scalability and ease of use of self-report assessments with the objective measuring of unguided responses possible in live assessments. The current research explores such an assessment methodology: reactive, open-response assessments (RORAs). RORAs consist of a set of virtual agents (i.e., virtual human characters) and environments with which a respondent interacts via unguided, free-text responding. In a given assessment, a specific skill or attribute can be objectively assessed based on a respondent's behaviors in a tailored, interactive scenario. Importantly, the inputs provided by a respondent in a RORA are unguided and unprompted. That is, respondents can elect to "talk" to agents at any time by starting to type what they would like to say. The RORA pauses during the respondent's typing (mitigating problems that might arise from differences in typing speeds) and uses natural language processing (NLP) algorithms to interpret the free-response text when the respondent is finished typing. The interpreted response is then used to drive the outcomes in the unfolding scenario. For example, if an agent is engaging in unprofessional gossip about members of his unit, and the respondent reprimands him, the scenario will continue in a different way than if the respondent encourages sharing more details. Not responding during an interaction (i.e., "remaining silent") can be interpreted as readily as any other response. If

an agent is carrying on a conversation with the respondent, it may pause for a brief period of time between statements as one would in a real-life conversation, but if the respondent elects to remain silent, the agent will carry on in whatever way makes sense given a lack of response. The flexibility of agent reactions presents an important challenge associated with developing RORAs. That is, there must be a set of scripted agent responses for any and all reasonable responses made. Another challenge involves the development of NLP algorithms that can properly interpret the potentially wide-ranging responses. The remainder of this paper describes an initial effort to address these two challenges in the context of developing a RORA for OCS, and to briefly describe the results of a validation effort seeking to determine if a relationship exists between the way respondents behaved in the RORAs and their interpersonal leadership skills as traditionally measured by OCS.

METHOD

Designing the RORA Scenarios

The purpose of a RORA is to assess a target population for a specific set of skills or attributes via realistic scenarios. Therefore, its development must take into account the characteristics of the target population to-be-assessed, the definition of the attribute or skill to-be-assessed, and the range of response options available in a given situation. These factors are considered in turn below.

Characteristics of the Target Population

For the current research, the focus was to develop a RORA for the interpersonal leadership skills of Officer Candidates (OCs) in Officer Candidate School (OCS). OCS trains individuals seeking to become Officers in the Army. The qualifications to enroll in OCS are: 1) being a US citizen, 2) having a 4-year college degree, 3) being between 19 and 32 years old, and 4) being eligible for a secret security clearance. While some OCs are prior Enlisted Soldiers, most have no prior military experience, so it was necessary that scenarios did not require advanced technical/tactical knowledge in order to comprehend events as they unfolded. One practical effect of this decision was that standard Army jargon could not be fully integrated into the scripts of scenarios because less authentic, but more readily understandable phrases would benefit novice audience comprehension. Another implication was that scenarios should put the OCs in a set of situations that might be faced very early in their careers. In light of this requirement, OCS instructors were interviewed to elicit examples of situations where they needed to use interpersonal leadership skills during their early days in their first unit. Scenario storyboards used these inputs wherever possible.

Defining the Skills To-Be-Assessed

In order to properly understand how Army leader interpersonal skills might be manifested in a real-world context, researchers collaborated with representatives from OCS. Recall that OCS uses the Army Requirements Model (FM 6-22) to define aspects of interpersonal leadership relevant to effective Officer performance. Following discussions about the contents of FM 6-22, the research team decided to focus the RORAs on a subset of skills which would be observable in a small window of time. Therefore, certain elements of FM 6-22 were not assessed in the current set of RORAs. Additionally, teasing apart skills that overlapped conceptually was not considered crucial for an initial test of the RORA methodology. This left five skills to assess within the initial RORAs: "Leads Others," "Develops Others," "Creates a Positive Environment," "Communicates," and "Gets Results" (see Table 1). These five skills became the foundation for storyboarding an initial set of scenarios for the RORAs. Three basic scenarios were developed. The first was based on a set of situations involving peer pressure or attempts at sharing gossip. These kinds of situations present challenges to professionalism, requiring leaders to actively "Lead Others" and "Create a Positive Environment." The next scenario was based on tensions between unit needs and the personal needs of the unit's members. Such situations were used to assess "Communicates" and "Gets Results." The final scenario was based on the idea that an existing unit to which a young leader was assigned may have developed bad habits or a toxic culture prior to the leader's arrival. This scenario constituted a promising opportunity for assessing "Develops Others" and "Creates a Positive Environment."

Determining Range of Response Options

RORAs are intended to elicit unguided behaviors from respondents; however, these behaviors must be anticipated by developers in order to program the agent behaviors that should happen in response to any given respondent behavior. That is, each scenario must be designed with "branches" that represent the course of events that unfolds given a

particular behavioral choice made by a respondent. See Figure 1 for a visual representation of the relationship between "scenes," "branches," "vignettes," and "scenarios" as those terms were used in this work.

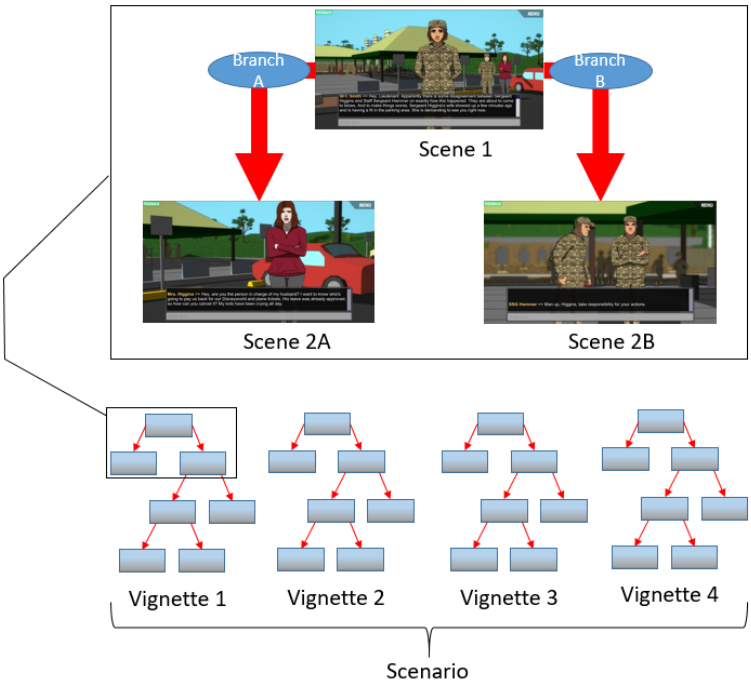

Figure 1. Structural Diagram of Scenarios

RORA Prototype. In order to determine how many branches would need to be developed for each scenario, a prototype incorporating a skeleton of each scenario was built in a PowerPoint Presentation (PPT). The PPTs were designed to allow respondents to input a response, if desired at any point, by clicking on an interactive "Talk" button while viewing the presentation. This design allowed for unprompted respondent input. While these presentations did include the storyline and the associated branching for each of the three scenarios that had been anticipated prior to respondent feedback, they did not include NLP to evaluate respondent behaviors. Instead, they required the use of a "controller" to direct branching based on respondent input. Presentations were designed such that after a respondent interrupted a scene and typed a response, they were presented with path options in the form of buttons (e.g., "A," "B," and "C") to select from. The controller would read a given response, and based on predetermined rules, would direct the respondent to click a button to lead them to the next most appropriate scene. PPTs were designed such that they captured all respondent input, the scene at which the input was given, and the branch taken to the next scene. Controllers were also able to capture feedback on storylines from respondents and instances in which additional branches were required. PPT prototypes were administered to Army junior leaders from three different units as depicted in Table 2. Variability in experience levels and backgrounds across the units was considered helpful for fleshing out a reasonable range of potential responses.

Table 2. Prototype Administration Participants

Unit	Participant Ranks	N
Infantry Brigade Combat Team (IBCT)	2LT, 1LT, CPT	63
Maneuver Captains Career Course (MCCC)	1LT, CPT	36
Officer Candidate School (OCS)	Officer Candidates	24

Through the use of the PPT prototypes, researchers updated storylines, added and refined branches within the scenarios, and collected user input to build libraries for use in the development of NLP algorithms. Special effort had to be taken to manage the number of reasonable branches deriving from each scene (especially if the scene happened early in a given scenario) in order to avoid an exponential growth in the number of scenes that would need to be programmed. In instances where too many branches were identified, clarification of specific details were typically

inserted in the narrative that would tend to drive respondents away from certain responses that did not serve to address the to-be-assessed skills. This technique typically resulted in a given scene having two or three reasonable branches. (Note that "reasonable" in this context does not indicate that all branches derived from appropriate leadership choices. Indeed, many of the branches in a given scenario would only be encountered if suboptimal leadership behaviors were demonstrated.) Further, scenarios were broken into multiple "vignettes" or sequences of related scenes. Each vignette typically consisted of a three to four scene sequence, meaning a given vignette would require programming roughly 12-15 scenes to account for all reasonable branches (see Figure 1). Between each vignette, a narrative section was inserted to keep the overall flow of the scenario intact regardless of the outcome of a given vignette. Again, this was done to reduce the total number of scenes that would need to be programmed for each scenario. By the end of the development process, scenarios ranged from 39 to 60 total scenes.

Natural Language Processing

Respondents interacted with virtual agents in each of the scenarios via free-text inputs. Each time an input was made, the RORA needed to determine which branch to follow in order to appropriately account for the most recent input. Responses to the PPT prototypes initially informed development of the NLP branching structure. To facilitate this requirement, a language processor had to be developed to decode respondent input and deduce its intent. The NLP algorithms developed for the RORA consisted of two distinct sub-systems: a parser and a keyword matcher.

Parser

The parser sub-system handled most of the NLP tasks. The parser was constructed with various components including a dictionary, a tokenization subroutine, a stop word removal subroutine, a spell checker, a lemmatization subroutine, a subroutine that tagged each word with a basic part of speech, and a subroutine that parsed all candidate choices available in the libraries specific to each scene allowing the keyword matcher to identify matches. The dictionary included nouns, verbs, adverbs, adjectives, and prepositions. To increase parser sensitivity, the dictionary also included variations/synonyms for each word, the lemma (i.e., root form of the word), a list of common misspellings, and a list of known expansions (e.g., "dunno" should be expanded to "don't know"). The sequence of parsing was to first "tokenize" the input, splitting any sentence into a list of distinct words and punctuation items (e.g., "I dont like all the running" would become "I, dont, like, all, the, running"). Next, "stop words" (e.g., a, an, the, of) were removed from the list. A spell check was then conducted (e.g., "I dont" became "I don't"). Then, "expansions" were applied (e.g., "I don't" became "I do not"). Lemmatization was then conducted in order to find the root word or base synonym for each word (e.g., "running" became "run"). Finally, all words were tagged with basic part-of-speech (e.g., I {pronoun}, do {verb}, not {adverb}, like {verb}, all {adjective}, run {noun}).

Keyword Matcher

After parsing the text, the resulting data was processed via a keyword matcher that generated a match score (i.e., the degree to which the most recent input shares commonality with each candidate choice for the current scene). Candidate choice libraries for each scene were composed of previously identified potential inputs that were paired with a specific branch to a new scene. The keyword matcher consisted of two subroutines: one to generate a match score for input-candidate pairs, and one to process those match scores. The generating subroutine determined match scores based upon several factors. First, it used the value of each of the input's words, where value is determined by how unique to a particular branching choice a word is (i.e., common words have a lower value). Next, it used the proportion of words that were matched in the candidate and the proportion of words that were matched in the input. Then, it used the position of each word in the input and the similarity between input and candidate words, where similarity was based first on part-of-speech and then lemma. The processing subroutine returned a "match" for an input that scored above a given threshold, then compared all "matches" and chose the highest scoring candidate resulting in the scenario branching to the scene associated with the selected candidate. If present, it chose a "catch-all choice" if no choice matched or it informed the respondent their input was not understood

The RORAs were designed to allow respondents three attempts for input to any given scene. After respondents made three attempts with no "match," the scenario proceeded to a new preprogrammed scene.

RORA Refinement Process

Using the information gathered from PPT prototype data collection efforts, software developers compiled an initial version of the RORAs using the UNITY engine. This version included updated scenarios, branching, enhanced

artwork, NLP algorithms, and the capability to capture, store, and report all respondent input. The initial version of the RORAs was tested with OCs. Following each data collection session, data were used in much the same manner as with the PPTs to further refine the branching rules and NLP.

Multiple versions of the RORAs were developed using information and feedback from data collections, as well as refinements made from research team testing. Later versions of the RORAs included a tutorial for respondents to learn how to interact with the virtual agents in the scenarios. The tutorial was added following user feedback indicating that some OCs did not realize how the interactions worked until a few scenes into the first scenario. These later versions of the RORAs had the OCs take the tutorial prior to accessing the first scenario.

Subsequent versions also included a "Feedback" capability for respondents to leave additional feedback on scenarios, programming issues, and any other suggestions as desired. An additional, hidden capability was embedded to allow researchers to use a "Cheat Menu" to jump to any scene in the scenario instantly. This capability proved to be a valuable tool in navigating around programming issues in early versions of the software.

RORA Testing

Participants
Officer Candidates from three companies –Company #1 (17 OCs), Company #2 (119 OCs), and Company #3 (118 OCs) – participated in data collection sessions. Specific demographics were not collected for OCs, but the majority (~80%) were males in their early 20s. OCs provided their names to allow researchers to later acquire OCS performance data to determine if RORA responses related to aspects of course performance. All 254 OCs completed three RORA scenarios and were included in the NLP matching analyses. Of the 254 OCs, data related to course performance was obtainable for approximately 150.

Materials
OCs used laptop computers to complete the three RORA scenarios in groups of up to 12. The three scenarios included: 1) a "Hand Receipt" scenario focusing on the "Lead Others" and "Create a Positive Environment" skills, 2) a "Firing Range" scenario focusing on the "Communicate" and "Get Results" skills, and 3) a "Motor Pool" scenario focusing on the "Develop Others" and "Create a Positive Environment" skills. Table 3 describes each scenario by the set of vignettes it included and the interpersonal skill on which the scenario focused. Laptops were set up around a large conference table in an OCS conference room. OCs wore headphones to listen to agent voices in addition to being able to see transcriptions of their statements on the screen. Figure 2 shows a typical view of the RORA interface.

Procedures
The research team conducted seven data collection sessions. Prior to each session, OCs were provided with informed consent documentation describing the nature of the research and the general design of the RORAs. A researcher also provided an oral overview of the RORA technology and the manner in which inputs were to be made. OCs were told that individual inputs were best kept to a length of one or two sentences and that any inputs that were not able to be interpreted successfully by the RORA would be added in subsequent versions of the tool. After signing consent forms, OCs completed a tutorial demonstrating the use of the RORA, then completed the three scenarios in sequence. All OCs completed the scenarios in the same order, as a loose storyline was weaved across them and agents (members of a fictional Army unit) appearing in one scenario often reappeared in later scenarios.

Data collection occurred over several evenings, and each subsequent session utilized updated versions of the RORA to alleviate any known issues with programming and to continually add to the NLP libraries. Individual OCs were allowed to provide data on only one occasion. Data collection event information is depicted in Table 4. Following the completion of all data collection events, OCS instructors provided course outcome data for each of the members of the participating companies.

Table 3. Scenario Descriptions

Scenario Name	Vignette Descriptions	Target Leadership Skills
Hand Receipt	Subordinate shares gossip about members of the unit Peer attempts to pressure respondent into signing an inventory form ("hand receipt") prematurely The Platoon Sergeant assures the respondent she will take responsibility for finding items missing from an inventory inspection A Staff Sergeant berates a Private and provokes him to start a fight	"Lead Others," "Create a Positive Environment"
Firing Range	A set of peers is reluctant to assist the respondent in completing his/her orders to conduct a training event A Staff Sergeant complains about his leave being canceled just before he and his family are about to take an expensive vacation they have been planning for months The Platoon Sergeant is frustrated because one of the members of the unit forgot to bring necessary supplies to the training event The medical unit required to be present for the training event is running late despite the Commander's assurance he would coordinate with them previously While en-route to a meeting, a safety violation is reported over the radio at the firing range where the training event is being conducted The spouse of the Staff Sergeant who's leave was cancelled shows up to complain about the vacation being ruined Two Sergeants argue over who is at fault for the safety violation	"Communicate," "Get Results"
Motor Pool	A Specialist is relaying bad news to the Platoon Sergeant, who is not taking it well The Specialist privately complains that the Platoon Sergeant is using her as a scapegoat The Platoon Sergeant verbally abuses a vehicle maintainer who has failed to fill out a form properly A second vehicle maintainer complains that he is unsure if an oil leak is sufficient grounds for marking a vehicle as unavailable for an upcoming exercise A third vehicle maintainer cannot recall the specific problems a vehicle was having during its last inspection and the details provided on the form are too vague	"Develop Others," "Create a Positive Environment"

Figure 2. The RORA User Interface

Table 4. OCS Data Collection Events

Unit	Assessment Tool Version	N
Company #1, OCS	Version 1.0.2/1.0.3	17
Company #2, OCS	Version 2.0.3	73
Company #2, OCS	Version 2.0.4	29
Company #2, OCS	Version 2.0.5	17
Company #3, OCS	Version 2.0.6	41
Company #3, OCS	Version 2.0.7	42
Company #3, OCS	Version 2.0.8	35

RESULTS and DISCUSSION

The focus of this effort was the development of a novel methodology (RORAs) for the systematic assessment of interpersonal leadership skills in the context of US Army Officer training. Key issues included ensuring NLP algorithms could accurately interpret responses without the need for voluminous amounts of training data, and ensuring agents could behave flexibly in response to inputs. Additionally, the validity of the RORAs for assessing interpersonal leadership skills was also of keen interest. Results related to each issue will be discussed in turn.

Natural Language Processing Match Rates

Match rates for the NLP algorithms were measured in the percentage of inputs that could be matched to an existing element of the candidate choice libraries associated with each scene and successfully branch to a new scene. During RORA development, changes in these percentages across versions were largely due to the synergistic effects of improvements to the software and additions to the libraries. With each data collection, an updated version of the tool was used in which coding errors were corrected, improvements were made to existing processes, and additional processes were introduced to streamline and improve functions within the program. Additionally, new versions included NLP library updates from respondent inputs. The RORA was designed to save all inputs, including NLP matches and non-matches. Non-matched data was subsequently used to update NLP libraries which, in subsequent versions, would result in a match when respondent inputs contained language similar to the language used to update the libraries. With software and NLP updates throughout data collection efforts, the percentage of matches generally increased from initial to final versions.

Match Results

Figure 3 shows match percentages by version and scenario. The initial version of the RORA was able to successfully match 40%-50% of respondent inputs. Improvements were seen for the next four versions, leading to match rates in the 60%-75% range, which stayed roughly constant in the final two versions. Note that the initial version had 17 respondents, the following four versions had a total of 160 respondents, and the final two versions had 77 respondents. Thus, there was a fairly rapid improvement up to 50% greater than baseline, which then plateaued.

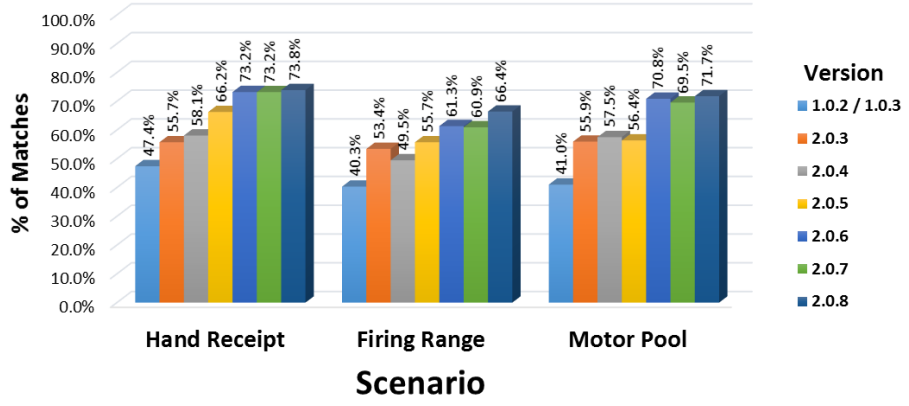

Figure 3. NLP Matches by Version and Scenario

Adjusting Match Rates

An automated report generator was developed to calculate match percentages. This tool provided output in the form of overall percentage of matches and non-matches for each scenario. Within each scenario, individual inputs were coded as either a match or a non-match for simplicity in report programming and analysis. In reality, there were many instances in which inputs were coded as a non-match when that coding unfairly penalized the NLP match rates. For example, the NLP has a routine that checks for misspelled words and can routinely account for common misspellings of words; however, typing skills and attention to detail in typing among respondents varied greatly. As a result, some misspellings were not interpreted by the software correctly, resulting in a non-match when otherwise a match would have occurred. Additional issues included: respondents using paragraph-length inputs despite advice to be concise, respondents using sarcasm, and respondents typing "gibberish," such as randomly pressing keys (presumably to skip past a given scene). Since it was not feasible within the bounds of this research to design the NLP to detect sarcasm, gibberish, gross misspellings, etc., a second "hand-scored" pass was made at evaluating match rates for the final version of the RORA. Match rates across the three scenarios improved between 2% and 4% after accounting for these factors.

Match Rate by Scene Response Frequency

Because respondents' inputs during a scenario drove which scenes they were exposed to, the amount of data available for incorporating into the NLP libraries differed from scene to scene. Scenes ranged from having as many as 246 responses (i.e., nearly all respondents supplied inputs) to having as few as 4 responses. An examination of NLP match rates as a function of the number of responses recorded shows that match rates for frequently visited scenes within a scenario climbed to nearly 90%. Table 5 shows NLP match rates for each scenario as a function of the number of recorded responses.

Table 5. NLP Match Rates for Scenes by Number of Responses

	Scenario		
Number of Responses	**Hand Receipt**	**Firing Range**	**Motor Pool**
201-250	87.2	74.1	83.3
151-200	88.5	69.2	72.6
101-150	85.3	70.1	78.2
51-100	73.4	56.5	73.5
1-50	52.1	63.8	70.9

These findings appear promising, as match rates reached high levels after collecting a modest amount of training data (less than 400 respondents across all data collection efforts) for the NLP algorithms. It is interesting to note that the "Firing Range" scenario produced consistently lower match rates than the other two scenarios. It is possible that the skills being assessed ("Communicates" and "Gets Results") required the crafting of scenes for which inputs needed to be more nuanced, diverse, complicated, or question-producing. As such, this additional level of input complexity would need to be accounted for when collecting NLP training data for future RORAs targeting those skills.

Agent Flexibility

One advantage of the RORA methodology is that inputs are unguided and unprompted, making the assessment of skills and attributes more like live assessments in their objectivity. This feature requires that agents are able to appropriately react to the range of inputs respondents provide. To determine if the RORA developed for this effort met this requirement, two metrics were used. The first was to ask respondents how adequately they felt the RORA flowed from one scene to another following the completion of all scenarios. 85.6% of respondents indicated that there were no recognized problems with the flow of scenes or reactions of the agents. Of those who did bring up problems, none mentioned more than one example of incongruous agent behavior.

A second metric for evaluating agent flexibility was to determine how many unrecognized inputs (i.e., those that did not lead to a previously defined branch) could be categorized as requiring additional branches. That is, some inputs that were not recognized could have logically led to previously defined branches if not for the word choices, misspellings, etc. of the respondents. However, some may not have been recognized because they were attempts to

drive the scenario in a direction that was not anticipated. Evaluating the unrecognized inputs for each scenario, 10.6% of unrecognized responses in the "Hand Receipt" scenario should have branched to unanticipated scenes. Accounting for these responses would have required the development of 6 new scenes. Analysis of the "Firing Range" scenario showed 7.3% of unrecognized responses were unanticipated and would require 7 new scenes. Finally, analysis of the "Motor Pool" scenario showed 12.4% of unrecognized responses were unanticipated and would require 6 new scenes. Given the number of total scenes in each scenario, this would suggest that agent behaviors accounted for between 80% and 90% of necessary behaviors to cover the range of reasonable respondent inputs.

Relationship between RORA Responses and OCS Performance

Given that the RORAs function at an acceptable level with respect to agent behaviors and NLP considerations, the issue of whether or not the RORAs appropriately assess meaningful differences in interpersonal skills remains. To determine if there is any relationship between RORA performance and interpersonal skill levels, an analysis was conducted using the scores OCs received on "Garrison Leadership" tasks in OCS and the scores they received on peer assessments. Of all the graded events in OCS, these two sets of scores are the most closely linked to the interpersonal leadership skills and constitute the most appropriate criterion data for RORA validation.

It became evident early in the analytic process that a restriction of range issue existed for the OCS performance data. Given this problem, the analysis was conducted using only those OCs whose average ratings on the interpersonal leadership skills were in the top or bottom 10 in the data set. The RORA performance for these 20 OCs were compared across each vignette. Mean scores and standard deviations on each vignette are shown in Table 6. Scores were determined by raters who reviewed the OC responses in each vignette. If an OC demonstrated the targeted skill within his/her set of vignette responses, the OC was credited with 1 point for the vignette. Thus, a mean score of 3.5 indicated that on any given vignette within a scenario, an average of 3.5 members of the group of 10 received credit for demonstrating the targeted skill. Student's T-tests revealed that group differences were not significant for "Hand Receipt" vignettes ($p<.44$), but were significant for "Firing Range" vignettes ($p<.00$) and marginally significant for "Motor Pool" vignettes ($p<.10$). These findings suggest a relationship between performance on the RORAs and traditionally measured indicators of interpersonal leadership skill, but that some vignettes were more sensitive than others.

Table 6. RORA Performance for Top and Bottom 10 OCs

Group	Scenario					
	Hand Receipt		**Firing Range**		**Motor Pool**	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Top 10	3.5	1.7	8.6	1.1	8.3	1.5
Bottom 10	3.3	2.6	5.6	1.2	6.3	2.5

Limitations and Future Work

The current research represented a first attempt to create a novel type of assessment tool blending the positive aspects of self-report and live assessment methodologies. While the results are promising, there are several improvements that could be made in future iterations of the RORAs. First, the current RORAs are linear in narrative. If an interaction with an agent could include multiple issues (like most real-life conversations/interactions do), it would be beneficial if the RORA could smoothly transition back and forth between issues in a nonlinear manner. This would increase the realism of the interactions and could allow for a more nuanced probing of the targeted skills/attributes. Second, agents currently have no "memory" from one scene to the next. Having respondent inputs impact the agents in a more lasting way would be beneficial. This would need to be accomplished in a way that does not exponentially expand the number of scenes necessary to complete the RORA. One potential approach would be to have individual scenes that accomplish the same goal (e.g., the Sergeant discusses why a vehicle is not working with a Specialist), but that use somewhat different dialogue as a function of tracked agent "states" representing things like mood or motivation. Current work is underway to examine the feasibility of such improvements. Moreover, some vignettes appeared to align more closely with traditional live measures of interpersonal leadership skills than others. Future work is required to determine the nature of the differences in those vignettes.

In conclusion, the findings of the current work alleviate two important concerns one might have with employing the RORA methodology. First, the challenges related to agent flexibility and NLP accuracy are not insurmountable. And, more importantly, the desired relationship between the outcomes of live assessments and the RORAs were clearly demonstrated. Thus, RORAs constitute a promising avenue for the scalable and objective assessment of interpersonal leadership skills and other intangible competencies.

REFERENCES

Bedwell, W. L., Fiore, S. M., & Salas, E. (2014). Developing the future workforce: An approach for integrating interpersonal skills into the MBA classroom. *Academy of Management Learning and Education*, 13(2), 171-186.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, 17(2), 245-260.

Headquarters, Department of the Army. (2015). *Leadership development* (FM 6-22). Retrieved from http://www.milsci.ucsb.edu/sites/secure.lsit.ucsb.edu.mili.d7/files/sitefiles/fm6_22.pdf

Klehe, U. C., Kleinmann, M., Hartstein, T., Melchers, K. G., Konig, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the idea employee factor. *Human Performance*, 25(4), 273-302.

Krumm, S., Lievens, F., Huffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? *Journal of Applied Psychology*, 100(2), 399-416.

Mumford, T. V., Campion, M. A., & Morgeson, F. P. (2007). The leadership skills strataplex: Leadership skill requirements across organizational level. *The Leadership Quarterly*, 18, 154-166.

Mumford, M. D., Zaccaro, S. J., Harding, F. D., Jacobs, T. O., & Fleishman, E. A. (2000). Leadership skills for a changing world: Solving complex social problems. *The Leadership Quarterly*, 11(1), 11-35.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, 15(3), 263-280.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2017). Applying organizational justice theory to admission into higher education: Admission from a student perspective. *International Journal of Selection and Assessment*, 25(1), 72-84.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. M. (1949). The American soldier: Adjustment during Army life. *The Annals of the American Academy of Political and Social Science*, 265(1), 173-175.

U.S. Army Combined Arms Center. (n.d.). *Developing leadership during unit training exercises*. Retrieved from https://usacac.army.mil/sites/default/files/documents/cal/DevelopingLeadership.pdf

The research described herein was sponsored by the Army Research Institute for the Behavioral and Social Sciences, Department of the Army (Contract No. W5J9CQ11D0001-0026). The views expressed in this paper are those of the author and do not reflect the official policy or position of the Department of the Army, DOD, or the U.S. Government.