Research Report 2030

Tools for Assessing and Tracking Leadership Attributes and Competencies

Victor J. Ingurgio
U.S. Army Research Institute

Krista L. Ratwani
Ben Nargi
Aptima, Inc.

Scott Flanagan
Frederick J. Diedrich
Sophia Speira

Tatiana H. Toumbeva
Aptima, Inc.

October 2020

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 ZBYLUT, Ph.D.
Director

Research accomplished under contract
for the Department of the Army by

Aptima, Inc.

Technical review by

Dr. Randy Brou, U.S. Army Research Institute

Dr. Ashley Wittig, U.S. Army Research Institute

NOTICES

DISTRIBUTION: This Research Report has been submitted to the Defense Information Technical Center (DTIC). Address correspondence concerning ARI reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPE-ARI-ZXM, 6000 6th Street Building 1464 / Mail Stop: 5610), Fort Belvoir, VA 22060-5610.

FINAL DISPOSITION: Destroy this Research Report when it is no longer needed. Do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Research Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
10-2020	Final	October 2018 – September 2019

4. TITLE AND SUBTITLE
Tools for Assessing and Tracking Leadership Attributes and Competencies

5a. CONTRACT NUMBER
GS00Q14OADU428

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
633007

6. AUTHOR(S)
Victor J. Ingurgio, Krista L. Ratwani, Ben Nargi, Scott Flanagan, Frederick J. Diedrich, Tatiana H. Toumbeva

5d. PROJECT NUMBER
A792

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral & Social Sciences
6000 6th Street (Building 1464/Mail Stop 5610),
Fort Belvoir, VA 22060-5610;

Aptima, Inc., 12 Gill St. Suite 1400, Woburn, MA 01801

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, Building 1464
Fort Belvoir, VA 22060

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

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Research Report 2030

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

13. SUPPLEMENTARY NOTES
ARI Research POC: Dr. Victor J. Ingurgio, Fort Benning Research Unit

14. ABSTRACT
The Army's People Strategy (U.S. Department of the Army, 2019a) enhances Army readiness by maximizing the talents of every person. Assessment is a key piece of ensuring the talent management strategy, as accurate assessment is the first step to providing Soldiers with the feedback and experiences they need to continue to develop themselves and to do their jobs effectively. To aid the Army in this initiative, a mobile application and accompanying dashboard were developed to allow leaders to better assess and track their Soldiers over time. The foundation of the application is facilitating accurate assessment of Soldiers on the attributes and competencies contained within the Army's Leadership Requirements Model (U.S. Department of the Army, 2012; 2015; 2019b). This report describes the development of the tools, as well as evaluations conducted to determine the usability/utility of the tool and inter-rater agreement and reliability.

15. SUBJECT TERMS
Talent Management, Leader Requirements Model, Mobile Applications, Performance Assessment

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. Jennifer S. Tucker
Unclassified	Unclassified	Unclassified	Unlimited Unclassified	50	19b. TELEPHONE NUMBER 706-545-2490

Research Report 2030

Tools for Assessing and Tracking Leadership Attributes and Competencies

Victor J. Ingurgio
U.S. Army Research Institute

Krista L. Ratwani
Ben Nargi
Aptima, Inc.

Scott Flanagan
Frederick J. Diedrich
Sophia Speira

Tatiana H. Toumbeva
Aptima, Inc.

Fort Benning Research Unit
Jennifer S. Tucker, Chief

October 2020

ACKNOWLEDGMENTS

We would like to thank all of the former and current Leaders and Soldiers at Ft. Campbell, KY who contributed to the development of these tools, in particular those of A TRP, 1-75 CAV, 2BCT, 101st ABN (AASLT). The creation of these tools would not have been possible without them.

TOOLS FOR ASSESSING AND TRACKING LEADERSHIP ATTRIBUTES AND COMPETENCIES

CONTENTS

TABLES

Page

LIST OF FIGURES

Tools for Assessing and Tracking Leadership Attributes and Competencies

Introduction

The Army's People Strategy (U.S. Department of the Army, 2019a) enhances Army readiness by maximizing the talents of every person. The Army's goal is to select, develop, and retain Soldiers who can respond effectively to various missions, the nature of which evolve over time. To achieve this goal, the knowledge, skills, and abilities (KSAs) of Soldiers must align with those missions. Hence, as described in the strategy document, assessment is a key piece of talent management. For example, within the Army's recent focus on Multi-Domain Operations (MDO; U.S. Department of the Army, 2018), Soldiers must be able to thrive in armed conflict as members of the Joint Force. Within this context, there is a need to "maximize human potential" (U.S. Department of the Army, 2018, p. 19) by carefully managing the development of Soldiers who can be successful in chaotic, dynamic environments. Therefore, it is essential to assess where Soldiers stand in relation to critical KSAs to determine if they are ready for specific missions and to determine where additional development may be needed.

One way that talent is managed in the Army is through the formal evaluation process that includes the completion of Officer and Noncommissioned Officer (NCO) Evaluation Reports (OERs and NCOERs, respectively). As part of the evaluation reports, Officers and NCOs are assessed according to their leadership attributes and competencies (Character, Presence, Intellect, Leads, Develops, and Achieves) detailed in the Army's Leadership Requirements Model (LRM; U.S. Department of the Army, 2012; 2015; 2019b--this version added the attribute of "Humility" to Character and is not assessed in this paper). It is challenging for leaders to assess performance and provide meaningful developmental feedback in that there may be various interpretations of the exact meaning of the attributes and competencies; in addition, it may also be difficult for leaders to accurately recall all of the behaviors exhibited by Officers and NCOs in relation to the attributes and competencies over time, especially as more negative and/or recent events will be more salient (Steiner & Rain, 1989). In addition to completing the OER/NCOER, leaders must also provide more formative feedback to Soldiers on an ongoing basis to help each Soldier better develop the leadership attributes and competencies. Providing such feedback has similar challenges, such as not capturing enough detailed information (both positive and negative ratings) to ensure that the feedback is effective.

To help leaders more effectively provide both formal and informal feedback to Soldiers on critical leadership attributes and competencies, assessment and tracking tools were developed. Specifically, the tools consist of a mobile application where leaders can quickly enter and track performance on the leadership attributes and competencies, and a back-end leader dashboard where information from across multiple mobile applications can be rolled up and displayed for a summary view of performance across a unit. While the tools were developed for use by Section Leaders to rate their junior Soldiers, they are generalizable to other ranks.

The foundational elements of the mobile application are short definitions of each leadership attribute and competency within the LRM as well as a five-point rating scale that provides behavioral examples of each attribute and competency at three anchors (1 = *Developmental Need*; 3 = *Standard*; and 5 = *Strength*). The development of these elements is described fully in Dein, Ingurgio, Ratwani, Diedrich and Flanagan (2019). The overall goal of

the definitions and rating anchors is to promote a greater understanding of performance gradients within each attribute and competency, and to increase consistency by standardizing the ratings of each attribute and competency.

This report describes the development of the mobile application and leader dashboard, as well as the evaluation of the mobile application. Specifically, concerning evaluation, data were collected in reference to the usability and utility of the tool as well as interrater agreement and reliability data. The latter focused on demonstrating that the developed behavioral anchors and rating scales promote consistency among raters. Screen shots of the developed tools are contained throughout the report.

Tool Development

The tools were developed through an iterative process (human-in-the-loop) of gathering user feedback and input and creating and modifying mock-ups and prototypes. The cornerstone of this process was the close collaboration between end users (NCOs) and a multi-disciplinary team of research psychologists, retired Army NCOs, user interface designers, and software engineers. For this work, we collaborated closely with a unit from Ft. Campbell, KY. While this unit, a Reconnaissance, Surveillance, and Target Acquisition (RSTA) unit, served as the main source of feedback throughout the development process, personnel from other units also provided feedback at times. Thus, while some of the content is specific to the RSTA context, other units could use the tools as a strong foundation with minimal modification.

Over the course of about eight months, we observed multiple different training events and held focus groups with active duty Soldiers. The purpose of the observations and focus groups was two-fold. First, as described in Dein et al. (2019), the observations provided information about how the leadership attributes and competencies manifested themselves across training events, ultimately leading to the development of the behaviorally based scales contained within the tools. Further, that article explicitly describes how the developed rubric is fully based on Army doctrine (LRM), and it describes the challenges, lessons learned, and procedures for validating the rubric and associated attributes (see Appendix A for the complete rubric). Second, the observations and focus groups helped to derive the tool requirements largely based on identifying the requirements of, and challenges faced by, leaders assessing Soldier performance.

Assessment Requirements and Challenges

Based on the observations and focus groups, several assessment challenges and requirements were identified. First, from initial discussions with senior leaders within the RSTA unit, they expressed that the tool should *not* directly populate official evaluation forms such as the NCOER given the formal nature of the evaluation process. Instead, it should provide a mechanism for leaders to make more informed observations (sometimes referred to as "evidence"), leading to improved and standardized feedback, and ultimately facilitating more precise, complete, and nuanced formal evaluation reports. The notion of having a tool that could be used for more formative feedback was validated through focus groups with NCOs. Second, one of the challenges identified by the team with respect to assessments and feedback is that the amount of feedback that Soldiers receive throughout their career is variable and very dependent on the leader providing that feedback. Therefore, any tool developed should help make annual assessments as well as periodic counseling and feedback easier to do, so that the frequency of using the tool increases simply because it is easy to use. Finally, through the observations and

focus groups, we noted that leaders often have difficulties remembering event specifics for all notable behaviors demonstrated by their Soldiers over time. Although some leaders have methods for tracking performance and accomplishments, not all do. Better methods are needed to help leaders, especially new leaders, compile both positive and negative observations on any given Soldier. Thus, our tools should allow leaders to easily track and compile behaviors in relation to specific leader attributes and competencies. Those accumulated data will not only allow leaders to complete NCOERs based on more focused and specific events, but also facilitate the delivery of more actionable, formative feedback.

Initial Tool Conceptualization and Key Features

During observations and focus groups, members of the design team took note of the discussions and the expressed requirements of the users, as well as the reasoning and workflows that emerged. This information then led to prototypes and the development of quick drafts of whiteboards and then computer prototyping software to envision the beginnings of a tool that could be used to take notes, and allow annotations of the LRM attributes and competencies. Several features were conceptualized and developed to allow leaders to quickly assess and track Soldier performance on the LRM attributes and competencies over time.

Note taking. Throughout the development process, a comparable metaphor that the team targeted was the common Army Greenbook; the de-facto repository for the ever-present note taking practiced by Soldiers of all ranks. The team wanted to take the flexibility and fluidity of the Greenbook and enable it with technologies afforded to the Army in the 21st century. Specifically, the assumption made by the team was that leaders making ratings or observations on a Soldier would have a continuous dialog about a particular Soldier. During observations and focus groups, the team noticed that some leaders would often write memos to themselves in their Greenbooks. On their own, these individual pieces of information might not constitute much of an assessment, but when aggregated, they could begin to form a broader picture of Soldier performance and development.

Therefore, the team created a simple note taking feature, that acted similarly to a text message thread seen on modern smartphones (e.g., if an individual sent text messages to himself over time, there eventually would be a threaded conversation, filled with date/time stamps). Thus, the mobile application was designed to mimic this interaction.

Attribute and competency tagging and ratings. The next stage of tool development was to determine how to insert the LRM attributes and competencies into the note-taking component of the application. Again, following a text message metaphor, users often add attachments such as photos, movies, or emojis. With this in mind, the team created a similar interaction, wherein a user could write text notes and attach relevant data to support those notes. A user can attach media in the form of videos or photos – taken at the exact moment that a note is made or attached from a user's photo library on their phone. In addition, the team introduced a tagging feature, which pulled up the comprehensive list of the LRM attributes and competencies, including all the sub-attributes and sub-competencies (e.g., not just Intellect [a competency], but Mental Agility [a sub-competency of Intellect]), as well as short definitions of each and associated rating scales (as described in Dein et al., 2019 and seen in Appendix A).

The initial rating scale was a simple binary thumbs-up/down approach. Given that the tool was not intended to populate official documentation directly, the binary thumbs-up/down

approach seemed sufficient to bring key pieces of information to the leader's attention quickly. In focus groups, NCOs appreciated the simplicity of a binary scale, but they also reported that the application needed to capture greater nuance in their observations to enable them to determine how best to prioritize their limited time/resources. To force a decision in a simple binary thumbs-up/down manner could produce a superficial approach to assessing the true complexity of Soldier performance. The NCOs also remarked that rather than making a formal LRM rating, they might sometimes prefer only to note the occurrence of a leadership attribute or competency.

Therefore, a simple 5-point Likert scale was used, with anchors for items at '1' (*Developmental Need*), '3' (*Standard*), and '5' (*Strength*) with '2' and '4' being intervening ratings. In addition, the application was designed such that users could tag an attribute or competency with a note, including the option to choose not to provide a rating (in this case, the rating would be listed as 'N/A'). The result of this request produced gradients in colors (red, amber, and green) associated with ratings in the application.

Roll-up views. One goal of the application was to easily aggregate data generated in a way that could help leaders analyze the data and then advise their subordinates. As a result, the tool includes a roll-up screen that takes into account the tags made by a rater on a Soldier and aggregates them into a list format that is sorted by the main, higher order attributes, with the capability to see ratio bars (visualize the data) that indicate the spread of ratings for an individual. These ratio bars can also be expanded to view the ratings on the sub-attributes. If more context for the rating is desired, users can tap on the rating and be taken to the note (with any media included) where they made the initial rating.

This roll up screen also mimics the order in which the NCOER and associated support forms ask for information from an evaluator. Therefore, the information in the application can be used to support leaders as they complete the ratings/forms.

Note and rating exporting. The application exports data into an easy-to-read .csv format (used by Microsoft Excel™), as well as a .docx format (used by Microsoft Word™). This feature enables Soldiers of all backgrounds and expertise levels to leverage the data/text in flexible ways. Because the data are not exported to a proprietary format, users can take the generated data, easily reformat it, email it, etc.

Leader dashboard. To supplement the data contained on individual mobile devices, a leader dashboard was also developed to allow leaders at higher echelons (e.g., company or battalion commanders) to view the ratings of a unit in an aggregated fashion. To develop the dashboard, we primarily relied on recommendations and feedback provided by leadership of the RSTA unit. Unit leadership indicated that they would not need to see notes/ratings associated with individual Soldiers. Instead, their primary interest was in getting a global view of performance at lower levels such as the squad/section, platoon, and even company/troop-level. They indicated that this information would be useful for training considerations.

For the leadership dashboard to be useful, we had to determine what visualizations of data were needed as well as how to pass data easily from the mobile applications to the dashboard. From a visualization perspective, positive feedback was received by users on two types of graphs: a bar graph (similar to the roll-up screen in the mobile application) to show the frequency of ratings across attributes and competencies, and a line graph to show changes in

attribute ratings over time. To synchronize the data, it was determined that a local area network with wireless routing would be most effective. In this case, the raters using the application would simply connect the mobile devices to a Wi-Fi network (provided by a supplied router). This Wi-Fi network connects the mobile device to a government imaged and approved laptop via a closed system that allows information to be passed from the mobile device to the laptop. The leadership dashboard then aggregates the data and displays the unit-level visualizations. As we developed the key features for this app, we were careful to protect Soldiers' identities by utilizing fictitious names.

Final Products

The final products consist of the Enlisted-Leader Attribute Requirements Application (E-LARA) that can run on both Apple's iOS and Google's Android operating systems, as well as an associated leader dashboard. The mobile application was developed to function on a tablet to enable use in the field for user convenience. The leader dashboard, which runs on a government imaged and approved laptop, was developed for use in the garrison environment. The workflow of each product is described below.

E-LARA Work Flow

The iOS and Android versions of the mobile application, with the exception of some minor aesthetic differences, operate within the same functional capabilities. To begin using the application, there is minimal initial set-up required. On its initial launch, the application requests access to a user's camera and photo library so that users can take photos and/or upload photos already on the mobile device. This feature will be further discussed below.

Entering Soldier information. To start, a user is presented with a landing screen, referred to as the *Home* screen (Figure 1).

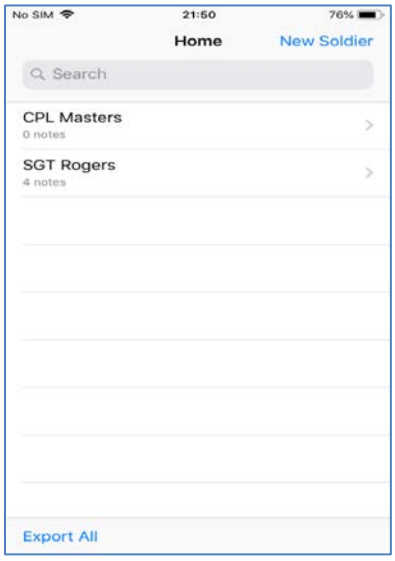

Figure 1. E-LARA Home screen with two Soldiers added. Fictitious names used.

Unlike other note-taking apps that tend to have a folder/file structure, E-LARA uses Soldier name/information as the default bucket in which notes are stored. For research, demonstration and test purposes of the app, the names we used were fictitious. However, if the app is approved for operational use, the Army would need to determine how this tool and its associated Leader Dashboard can be used while safeguarding a Soldier's Personal Identifying Information (PII) and data. A user then selects *New Soldier*, which initiates a creation dialogue (Figure 2). Since this is a closed system, users can enter a Soldier's Personal Identifying Information (PII) such as name (or roster number/DODID), rank, squad/section, platoon, and/or company/troop as appropriate. With the exception of a name, all other data are optional.[1]

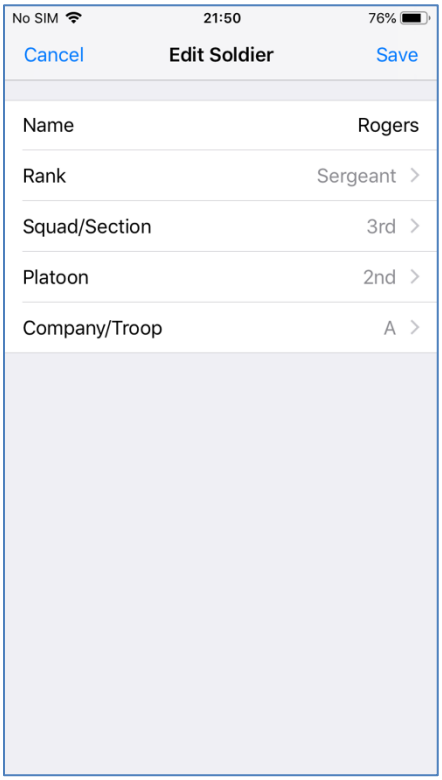

Figure 2. Create/edit Soldier screen; only a name is required (numbers/nicknames are encouraged to protect Soldiers' PII). Fictitious name used.

Upon entering in the desired Soldier information and selecting *Save,* users are taken to the Soldier screen (Figure 3). At first, this screen will be empty, however, within this screen, users can edit Soldier details (name, etc. through the *Edit Soldier* button), search through any notes (by text), sort the notes by date taken or by attributes present. At the bottom of this screen, users can export data in the form of .csv or .docx formats, view the Roll-Up (detailed later), or create a New Note.

[1] While the other data are optional, for the purposes of creating accurate roll-ups in the dashboard, it is a best practice to enter the unit information.

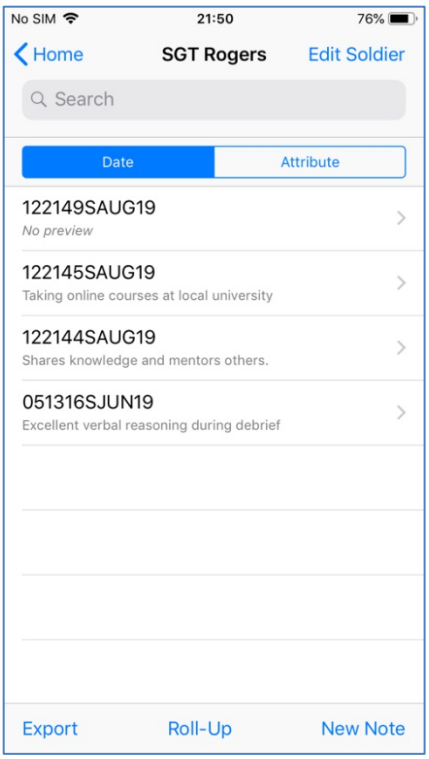

Figure 3. Inside a Soldier folder with four separate notes. Fictitious name used.

Creating a new note. Notes are the primary data component generated by E-LARA. Upon selecting *New Note*, users can input three types of data: Media, Tags, and Text (Figure 4).

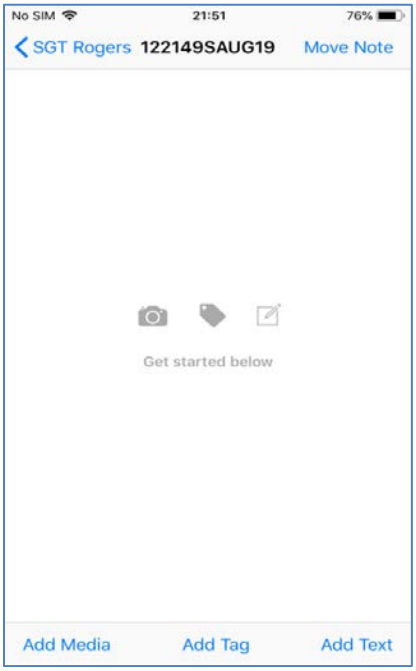

Figure 4. An empty note screen, with available actions (Add Media, Tag, Text) at the bottom. Fictitious name used.

Add Media provides the ability to add photos/videos either taken on the spot or selected from the device's photo gallery. *Add Tag* pulls up the Leader Attributes (detailed later), and *Add Text* pulls up a simple text input screen. Figure 5 shows an example for adding free text. Users are not constrained when adding text and it can be submitted in whatever format is chosen by the user (e.g., paragraphs). There is no character limit, and the text can be edited or deleted later.

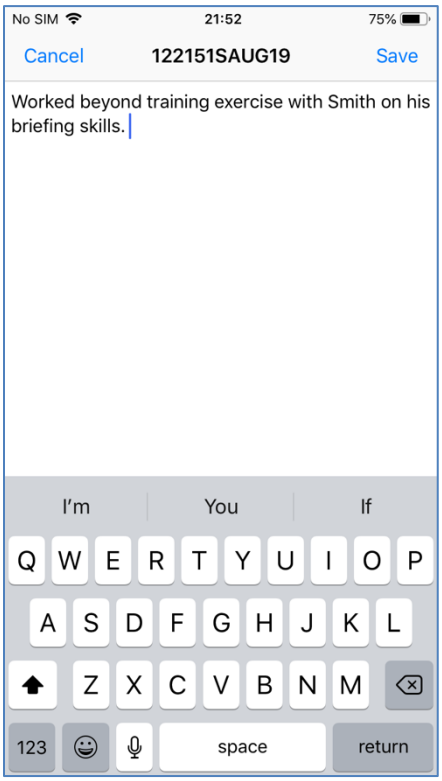

Figure 5. Entry screen for free text notes.

Adding a tag is one of the primary features of the application. Upon selecting *Add Tag*, a user can view a list of the 23 LRM sub-attributes and competencies grouped according to the related primary attribute/competency (Figure 6).

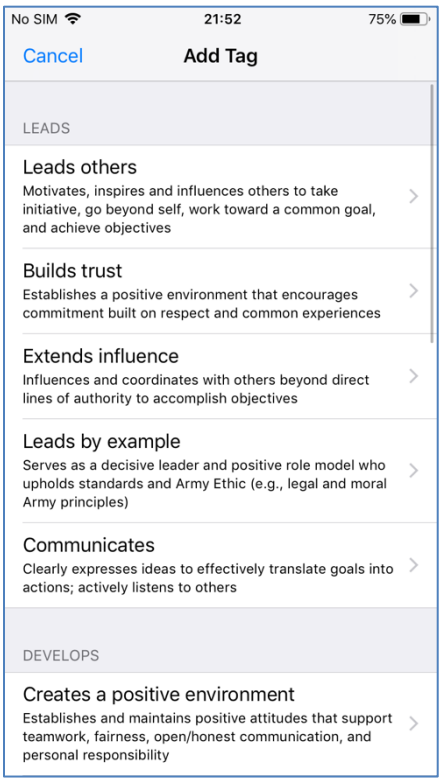

Figure 6. Scrollable list of LRM attributes and competencies, with a brief definition of each.

Once users identify an attribute they want to tag, they select it, and are then brought to a detailed screen that provides a short definition of the attribute, the 5-point rating scale, and the behavioral anchors corresponding to rating points of '1,' '3,' and '5' on the scale (Figure 7). From there, a user can simply choose to add the tag without a rating, or select a rating and then add the tag. If a user does not select a rating, but adds the tag, the tag will show up as "Gray" (no red, amber, green will be associated) and be considered as a Not Applicable (N/A) rating. A user can always return and either add/change/remove ratings from tags (and notes). Users may also add as many tags as they like to a note.

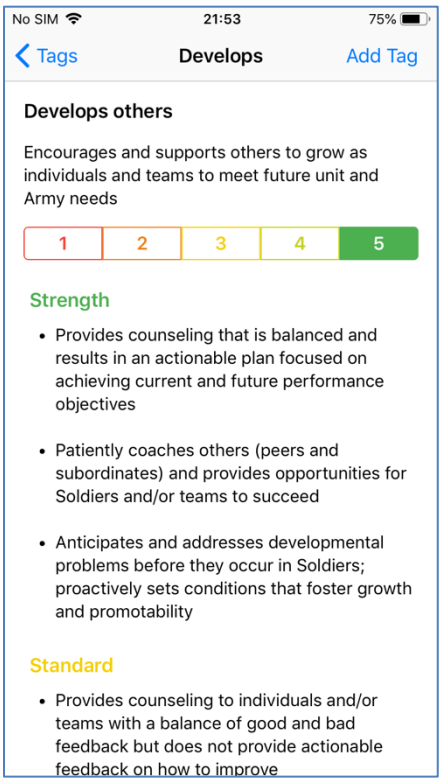

Figure 7. A selected competency, with the ability to add a numeric rating (1-5) based on the provided behavioral anchors.

Next, a user can add a photo or video using the *Add Media* button. This pulls up the standard smart phone dialogues that would typically be found in texting apps, messenger apps, etc. (Figure 8).

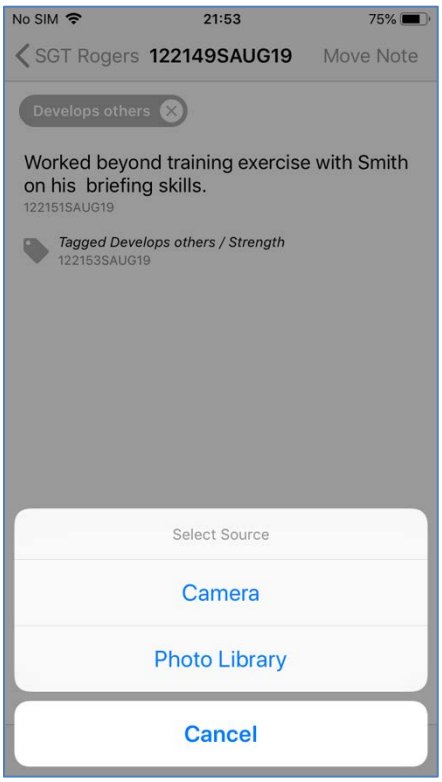

Figure 8. Adding media using the native sharing features in iOS (Camera/Photo Library). Fictitious name used.

Viewing notes and ratings. All notes taken for a given Soldier can be viewed in chronological order (newest to oldest) on that Soldier's home screen. Each note is date and time stamped to provide additional context around when the note was taken. Within the Soldier's home screen, users can utilize the search function to find specific words/phrases to help them filter down to the relevant notes. Users can also sort by attribute, which will pull up a list of all attributes tagged, while displaying the note in which they are referenced (Figure 9).

Data roll-up. The bottom of each Soldier's home screen contains a *Roll-up* button. When users select this button, they are taken to a screen that compresses all the tags and ratings made on that given Soldier into the six primary attribute and competency categories (Character, Presence, Intellect, Leads, Develops, and Achieves; Figure 10). Users can see the total number of tags for each of these categories, along with an average rating, plus a visual representation (known as the "Spread") that shows the frequency of each rating. This feature provides users with the overall profile of a given Soldier across all the LRM attributes and competencies. By selecting an attribute, the sub-attributes are then displayed along with the note that contains the attribute. From there, a user can decide to select the note to get more detail, or continue to explore other attributes. Within the Roll-up screen, any media are also compiled into an easily accessed grid wherein users do not have to jump from note to note to find relevant media they may want to reference. Finally, as a reference within the Roll-up screen, a user can view the complete set of short definitions developed for each leadership attribute and competency.

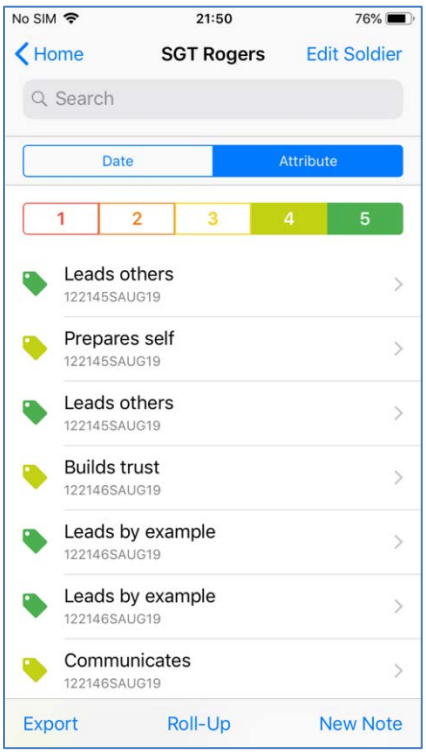

Figure 9. A list of attributes tagged to a specific Soldier, showing only those rated as a '4' or '5.' Fictitious name used.

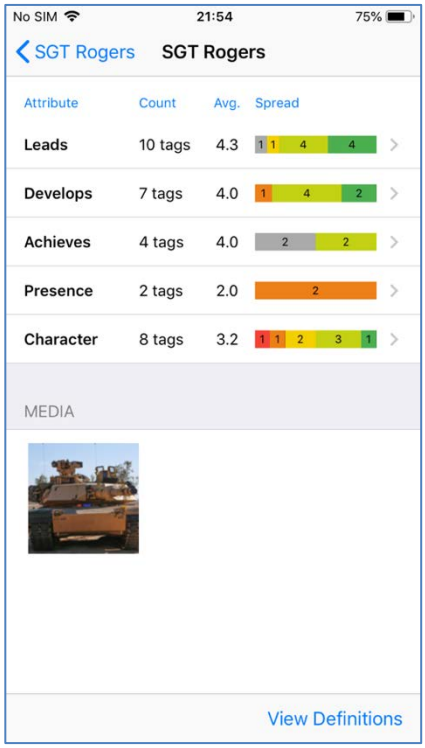

Figure 10. Viewing a Soldier's roll-up across all notes/ratings. Users can tap on an attribute to dive deeper into each note associated with the attribute. Fictitious name used.

The roll-up view is also available via the Home screen in E-LARA. In this view, all of the attribute and competency ratings made across Soldiers are compiled into a graph for one overall snapshot (Figure 11). This graph is for viewing purposes only. A user cannot drill down further, but rather it provides a simple visual summary of all ratings made across all Soldiers.

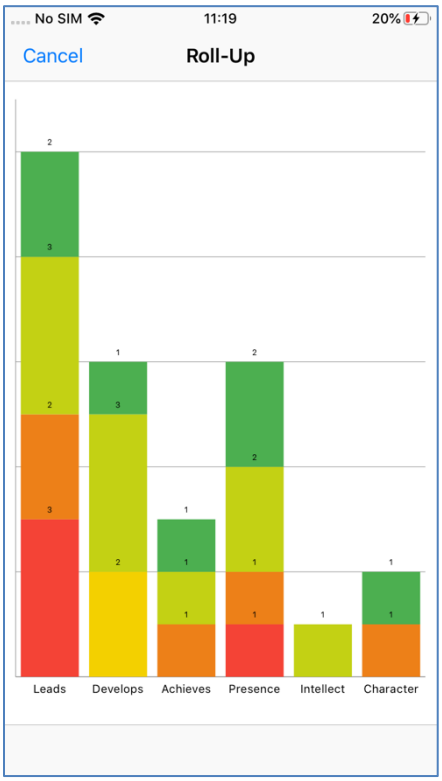

Figure 11. Group roll-up view that shows attributes and competency ratings (as well as frequency) across all Soldiers in the application.

Exporting data. E-LARA affords an easy way to export data for further analysis/editing using widely available programs such as Microsoft Word™ and Excel™. To export data, a user simply selects *Export* from the bottom of a Soldier's home screen or the app home screen (depending upon if the user wishes to export data associated with only one Soldier or with all Soldiers), and a share menu appears from which a user can email, text, or AirDrop a .csv or .docx file (Figure 12). The .csv has a user-friendly layout, and can be graphed/transformed using standard Microsoft Excel™ formulas and features.

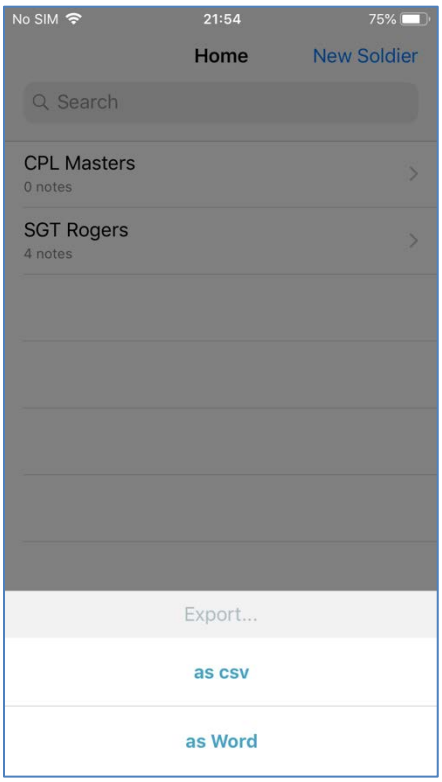

Figure 12. Exporting all data to either .csv or .docx file types. Fictitious names used.

Leader Dashboard

While the mobile application allows for the exporting of data from one mobile device, a desktop-based Leader Dashboard provides the ability to aggregate ratings from multiple mobile devices with several methods of data visualization. To synchronize data from the mobile application to the Leader Dashboard, a user with the application on a device must be within range of the Leader Dashboard's wireless network – from there a user would connect to the network, as they would any other Wi-Fi network. Once connected, a user selects *Sync* on the mobile device (found on the main app home screen), which establishes a local connection to the database. In less than a minute, the data on the application should transfer to the Leader Dashboard. We used a closed system for testing purposes. This process is depicted in Figure 13.

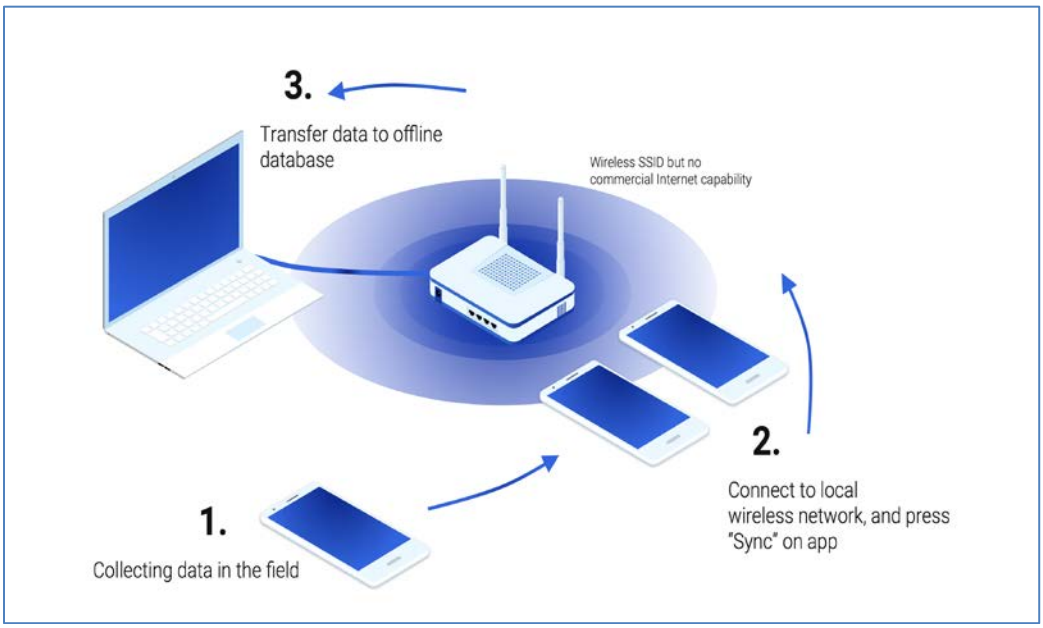

Figure 13. A depiction of the syncing process between the mobile application and the Leader Dashboard.

Within the Leader Dashboard, users can select different echelons by which to view the Soldier data (e.g., section, platoon). During this procedure, PII and free text notes associated with each individual's data are concealed to preserve confidentiality between Soldier and rater. However, it does allow a leader to see all the ratings made on that individual and a unit collectively. It is anticipated that the Leader Dashboard will be a resource that fosters the growth of offline conversations with Soldiers, raters, and leaders. Figures 14 and 15 display data visualizations present in the Leader Dashboard. Users can quickly select different visualizations and print these to share with others.

Figure 14. Leader Dashboard showing compiled ratings across Sections A and C as a diverging stacked bar chart.

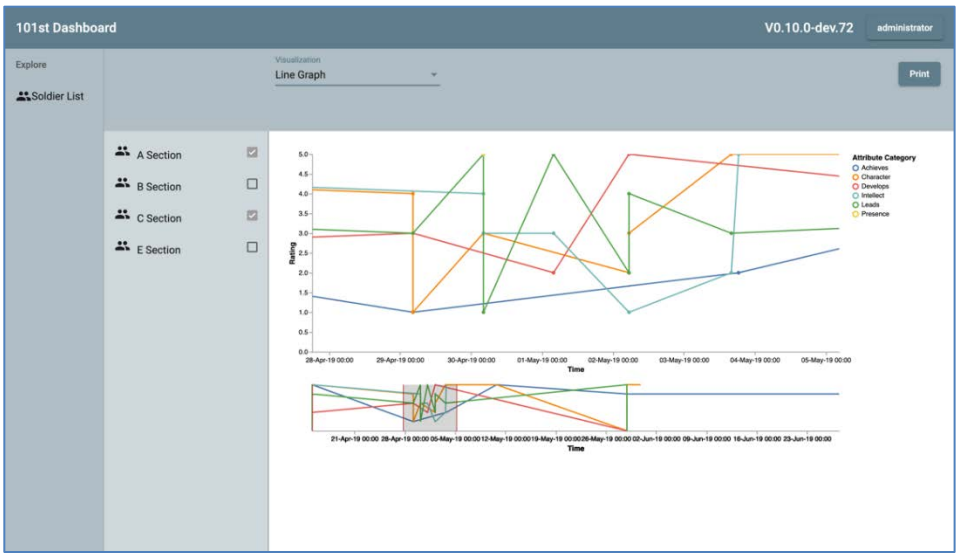

Figure 15. Leader Dashboard showing compiled ratings across Sections A and C as a time-based line-graph.

Usability and Utility Testing

The research team introduced a series of metrics and questionnaires to begin the capturing and analysis of usability and utility data. Usability is defined as the level of satisfaction one gets by using the application, including the efficiency of use (e.g., easy to use, free of errors, minimal effort to learn), and the effectiveness of using the application. Utility is defined as the usefulness of the application (e.g., the application does what it is supposed to do). Usability is often subjective and is measured by conducting interviews, usability studies, and surveys. Utility can be objectively measured, as this is the application's ability to perform measurable tasks and/or commands.

The research team's goal was to design a tool that had a high degree of utility, but also provided a deep degree of usability, such that users would gain satisfaction in using the tool while still accomplishing their professional tasks and goals.

Measures

In order to establish a baseline understanding of usability and utility, the team implemented the usage of the System Usability Scale (SUS; Brooke, 2013) along with a few free-response questions, as well as a Yes/No recommendation question (see Appendix B for the full questionnaire used). The acquisition of the SUS scores will allow the team to have concrete benchmarks to measure against for future endeavors. The free-response questions provide the research team with ideas for new features/uses, and the recommendation question provides a straightforward response as to whether a participant would be willing to use the application in the future.

Participants respond to the 10 items on the SUS on a five-point Likert style scale, i.e., '1' (*Strongly Disagree)* to '5' (*Strongly Agree).* Once the data are recorded, they are transformed and converted into a final SUS score, with a range between 1 and 100. Rather than being a straight percentage (i.e., a SUS score of 100 being twice as good as a SUS score of 50), a SUS

score is more analogous to letter grading conventions found in traditional educational settings (Bangor, Kortum, & Miller, 2008). There have been multiple, large-scale studies to help determine where a SUS score falls from a semantic understanding perspective (Bangor et al., 2008). As shown in Figure 16, a SUS score ranges between 0 and 100, with a score of 0 to 20 obtaining a *Worst Imaginable* rating, whereas a score of about 52-73 yields an *Ok* to *Good* rating with regard to usability; on the high end of the spectrum, a score of 100 is equated with a rating of *Best Imaginable*. These ratings can be reasonably mapped to a traditional grading scale (e.g., 70-79 is a C grade, 80-89 is a B grade). Bangor et al. also map the scores to an Acceptability Range, which is defined as being acceptable for field usage. Acceptability ranges from *Not Acceptable*, to *Marginal*, to *Acceptable*. According to Bangor et al., SUS scores below 50 indicate a cause for concern, suggesting that an application is unacceptable for field use. Scores in the 70s and 80s are considered promising, but do not guarantee a high degree of acceptability in the field.

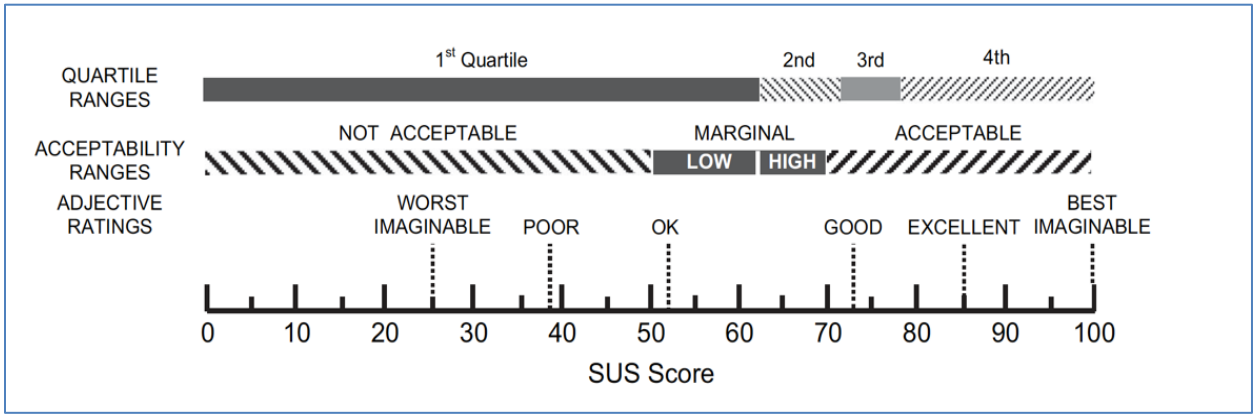

Figure 16. SUS score comparison chart adapted from Bangor et al. (2008).

Participants and Methods

The research team collected data from five participants who had used E-LARA over the course of a 30-day span. The participants included the troop commander (a Captain) and four team leads (all Sergeants). The participants had access to the application on both their personal smartphones, as well as government-furnished tablets. If they chose to utilize their own devices, either iOS or Android were provided for use; the tablets ran on Apple's iOS. To protect PII, participants were directed to use nicknames for the Soldiers that they were tracking. Further, the ratings that the participants made were part of this pilot (i.e., the rubric and tool was not used in place of formal performance ratings), and their ratings applied only for the pilot testing of this tool and did not get recorded or used in any unit performance records. Thus, the participants rated the Soldiers' performance only as exemplars for testing the capabilities of this tool.

Prior to the testing period, participants were given an in-person brief to demonstrate the features, usage, and anticipated workflow of the application. Beyond this brief, no other instructions or support were provided. The research team returned after the 30-day period and administered the measures described above and in Appendix B. A brief focus group was also conducted about the use of the application.

Results

The average SUS score was 88.5 across the five participants, resulting in an Adjective rating of *Excellent* and an Acceptability rating of *Acceptable* (see Figure 16). Three of the five participants said they would recommend E-LARA to members of their unit and/or colleagues; two participants would not recommend based on their own preferences for using a Greenbook. A summary of the survey data is presented in Table 1.

For the open-ended comments, as far as removing any features, none of the participants recommended the removal of any feature. Instead, participants suggested new features such as fillable forms (i.e. export to an NCOER), the ability to change or select different attributes for lower enlisted Soldiers, and adding more metrics such as physical training (PT) scores. The only change request for a current feature is the ability to change the date of a note (currently the notes are timestamped based on when they were created).

Table 1

Usability and Utility Results for E-LARA

Participant	SUS Score	Adjective Ratings	Acceptability Range	Recommend to Others?
1	80.0	Good	Acceptable	Yes
2	87.5	Excellent	Acceptable	Yes
3	100.0	Best Imaginable	Acceptable	No
4	77.5	Good	Acceptable	No
5	97.5	Excellent	Acceptable	Yes

Interrater Agreement and Reliability Data

Ultimately, the intended use of the E-LARA tool is to develop a pool of ratings for each Soldier across multiple months and multiple events as they perform their jobs in an operational unit. While utility and usability feedback was obtained from the intended users in the operational setting, data related to rater agreement could not be obtained due to only one leader rating each Soldier. Given the structure of the units and demands on time/operational tempo, it was not possible to assign multiple raters to individual Soldiers.

Consequently, to explore issues related to percent agreement and interrater reliability, the research team relied on two other settings to collect data from multiple raters evaluating the same events and Soldiers. First, members of the research team employed the measures while

observing a Field Training Exercise (FTX). The purpose of this initial data collection was to develop an understanding of factors that influence how different observers employ the E-LARA tool in an event, thereby shedding light on issues that might affect agreement. Second, use of the tool was examined more extensively with a large group of Soldiers in the context of peer evaluations. This latter case had the advantage of exploring interrater agreement and reliability based on mutual experiences acquired over the course of a few months, albeit from the perspective of peers rather than leaders. In addition, because the Soldiers were junior, they had limited experience with the LRM. Taken together, these data are suggestive of, but not specific to, the ultimate intended operational use setting.

Field Training Exercise

The initial use of the E-LARA tool to examine agreement between raters focused on an FTX within an institutional training setting. More specifically, members of the research team employed E-LARA while observing Soldiers executing an exercise that included elements such as Call for Fire, React to Contact, Casualty Evacuation, and Key Leader Engagements. The FTX therefore provided a rich set of events that could potentially involve many of the attributes and competencies in the LRM.

Within the context of this FTX, raters employed E-LARA over approximately eight hours of observation over the course of two days. Eight Soldiers were observed during this time by pairs of observers for approximately two hours each, although in one instance three individuals conducted observations of one Soldier, resulting in 10 sets of paired ratings between observers. The raters included five members of the research team operating in various combinations across the two days. To ensure confidentiality of all ratings, raters were directed to identify Soldiers with a single digit numeric designator. The scores were only recorded as part of the pilot testing of the capabilities of the tool, and the scores were not recorded as part of any formal unit training process or evaluation.

FTX rater agreement results. Overall, across the two days of observation, there were 60 pairs of ratings available to explore interrater agreement (i.e., comparing ratings across rater pairs on the same attribute/competency for the same Soldier). Across pairs, percent agreement, as calculated by the paired ratings being within one point of each other, was 80%. More specifically, 38.3% of ratings were identical, and 41.7% differed by one point (with 15% differing by two points, and 5% differing by three points, with no instances of differences of four points). These results suggest that raters generally agreed in terms of performance being relatively high, low, or average such that, for example, it was rare for one rater to rate an attribute as average whereas another rated the same as superior or poor. That being said, this exercise also uncovered another challenge associated with use of the E-LARA tool. In total, there were 60 pairs of ratings, and another 91 cases in which neither observer rated an attribute; however, there were another 79 instances in which only one observer rated an attribute for a Soldier who was not also rated by a second observer. In other words, in terms of agreement on what to rate (or not), the raters agreed on 65.7% of cases [(60+91)/230]. Even though the raters were seeing many things similarly, there were cases in which they focused on different aspects of performance amidst the complexity of the FTX.

To explore these issues further, data were split between observation days (see Table 2). The two days were separated by a session in which raters discussed differences in what they observed in an effort to synchronize their rating methods regarding the kinds of behaviors that

related to the attributes and what constituted various levels of performance. This rater "training" session ended up being essential, as overall interrater agreement within one point increased from 71.8% on the first day to 95.2% on the second day. In addition, there was also a difference in the percent of cases in which only one rater rated a particular attribute. On day one, there was 63.0% agreement on what to rate, and day two agreement on what to rate increased to 69.6%. These data suggest that the rater training helped both in terms of ratings within an attribute as well as what attributes raters chose to rate.

Table 2

Agreement Data Broken Up across Rating Days

Measure	*Day 1*	*Day 2*	*Overall (Day 1 and Day 2)*
Interrater Agreement for Paired Ratings[1]			
Identical	23.1%	66.7%	38.3%
Within 1 Point	48.7%	28.6%	41.7%
Within 2 Points	20.5%	4.8%	15.0%
Within 3 Points	7.7%	0.0%	5.0%
Within 4 Points	0.0%	0.0%	0.0%
Agreement on What Attributes Rated			
Number of Paired Ratings	39	21	60
Number of Non-Paired Ratings	51	28	79
Number Rated by Neither Observer	48	43	91
Percent Agreement[2]	63.0%	69.6%	65.7%

[1]*Paired ratings* examined the agreement across raters when both raters provided a rating on the same attribute.
[2]This *percent agreement* accounts for overall agreement across paired ratings and cases in which neither rater scored an attribute, divided by the number of possible paired ratings (sum of paired ratings, those rated by neither rater, and those rated by only one rater).

Summed across days, the team also examined agreement for individual attributes. Given the 60 total pairs of ratings across 23 attributes, these data are very sparse and must be interpreted with caution. That being said, a few notable trends emerged. First, there were no paired ratings for the attributes of Extends Influence, Fitness, Innovation, Prepares Self, Stewards the Profession, and Warrior Ethos. While the FTX did not set conditions for some of these attributes to be observed (e.g., Fitness), it is likely that others could have been observed but were not the focus of attention (e.g., Prepares Self). Second, in contrast, the highest number of paired ratings were for Sound Judgment, Communicates, and Confidence, suggesting that these attributes and competencies were salient given the nature of the FTX. These three items also had the most paired ratings with differences within a single point. Third, it was not the case that an attribute or two accounted for most of the larger difference between raters (i.e., those of two or three points). Rather, the three instances of differences of three points were spread across three attributes, and the nine instances of differences of two points were spread across eight attributes. Builds Trust had two instances of differences of two points, but it also had three instances of no difference and one instance of a difference of a single point. Even when rolled up to the higher-level categories in the LRM (i.e., Character, Presence, Intellect, Leads, Develops, and Achieves), the three instances of three point differences were in different higher-level categories (one each in Character, Presence, and Intellect). For the two point differences, they were distributed across

Character (2), Presence (2), Intellect (1), and Leads (4). Develops and Achieves had no instances of two- or three-point differences. When viewed as a percentage of instances of all paired ratings, instances of differences of two points or more were 37.5% for Character, 18.2% for Presence, 14.3% for Intellect, and 26.7% for Leads. This analysis suggests that Character and Leads may be more challenging to rate consistently.

Collectively, these data point to a few insights. Most importantly, the between-days data suggest that discussions between raters as part of "training" can help improve agreement on what to look for and how to evaluate it. While use in the FTX was somewhat artificial, further discussions of how attributes manifest across contexts are likely to be especially valuable. These data also suggest that not all attributes will be salient in all settings, highlighting the need to collect data over time in multiple events in order to capture a more complete picture of an individual. Finally, there is little evidence that certain attributes are more or less problematic, although initial data point to the overall categories of Character and Leads perhaps being slightly more challenging in terms of obtaining consistent ratings. While these data are preliminary, rater agreement was explored more fully in the context of a peer evaluation process, described in the next section.

Peer Evaluation Exercise

The second use of the E-LARA tool was by a larger group of Soldiers in the context of peer evaluation at four different Army installations. The Soldiers, all assigned to their units within the last six months following completion of Initial Entry Training (IET), completed the peer evaluation process as part of another program [for additional details, see Tucker et al. (2019)]. Here, de-identified data were used to examine how Soldiers rated their peers using the E-LARA application across all the LRM attributes and competencies. Participating Soldiers first had a brief introduction to the application and embedded rating scales, and then used the tool to rate each of their peers on each sub-attribute and competency within the LRM. This exercise resulted in ratings by two to five peers for 90 Soldiers, resulting in 5,325 total paired ratings. In 655 cases, only one Soldier in a pair rated an attribute for a given peer. While the Soldiers were instructed to rate their peers on all items, some missed a few, accounting for these unmatched pairs. Possible reasons for missing items include confusion over directions and/or which items had been rated.

Peer evaluation interrater agreement and reliability results. Given the 5,325 paired ratings, percent agreement, as calculated by the ratings being within one point of each other, was 75.6%. More specifically, 31.4% of ratings were identical, 44.2% differed by one (1) point, 19.7% differed by two (2) points, 3.4% differed by three (3) points, and 1.4% differed by four (4) points. Overall, these percentages suggest that raters generally agreed in terms of performance being relatively high, low, or average. This general level of interrater agreement is remarkable given the very limited familiarity of the Soldiers with E-LARA and the LRM.

Building on these analyses, interrater agreement and reliability were examined at a more granular level. Specifically, percent agreement and intraclass correlation coefficients (ICCs) were calculated for each LRM sub-attribute and competency (see Tables 3 and 4, respectively).

As shown in Table 3, the highest percent agreement among raters in terms of providing identical ratings was found for Prepares Self (37.0%) followed by Mental Agility (36.6%), and Expertise (35.6%). When rolled up to the higher-order conceptual categories depicted in the

LRM, the highest percent agreement of identical ratings was found for Achieves followed by Intellect and Develops. The lowest agreement was found for Character and Leads.

At least 70% rater agreement was demonstrated for each attribute/competency when the provided ratings were within one point on the five-point scale. The highest percent agreement within one point was for Expertise followed by Communicates and Military and Professional Bearing. When rolled up to the higher-level categories, the highest percent agreement within one point was found for Presence followed by Intellect and Develops. The lowest agreement within one point was found for Achieves, followed by Character and Leads.

Table 3

Descriptive Statistics and Percent Agreement by Attribute/Competency

Attribute/Competency	***N***	***M (SD)***	**Average % Agreement across Pairs of Raters**	
			Identical Ratings	*Ratings within 1 Point*
Character	***891***	***3.7 (.8)***	***27.6***	***74.3***
Army Values	220	3.8 (.8)	24.2	79.1
Empathy	220	3.6 (.9)	28.2	76.6
Warrior Ethos	214	3.7 (.8)	26.1	73.7
Discipline	237	3.7 (.9)	31.7	72.8
Presence	***922***	***3.8 (.7)***	***30.5***	***80.8***
Military and Professional Bearing	231	3.7 (.8)	30.7	83.4
Fitness	241	3.8 (1.0)	32.8	74.6
Confidence	228	3.9 (.9)	29.5	83.0
Resilience	222	3.9 (.8)	28.7	82.2
Intellect	***1,052***	***3.6 (.7)***	***32.7***	***80.3***
Mental Agility	225	3.7 (.8)	36.6	79.1
Sound Judgment	221	3.6 (.8)	23.1	80.5
Innovation	193	3.5 (.8)	34.2	80.7
Interpersonal Tact	205	3.7 (.9)	33.8	75.9
Expertise	208	3.7 (.9)	35.6	85.1
Leads	***1,271***	***3.5 (.8)***	***26.1***	***75.3***
Leads Others	259	3.5 (.9)	26.2	70.4
Builds Trust	256	3.7 (.9)	27.7	78.1
Extends Influence Beyond the Chain of Command	254	3.5 (.8)	26.2	74.2
Leads by Example	247	3.4 (.9)	24.2	70.1
Communicates	255	3.6 (.9)	26.0	83.5
Develops	***941***	***3.5 (.8)***	***32.9***	***78.4***
Creates a Positive Environment	245	3.7 (.9)	31.8	73.1
Prepares Self	231	3.7 (.9)	37.0	78.1
Develops Others	239	3.5 (.8)	31.1	79.3
Stewards the Profession	226	3.6 (.9)	31.8	83.2
Achieves	***248***	***3.8 (.9)***	***33.3***	***71.9***
Gets Results	248	3.8 (.9)	33.3	71.9

Note. *N* in the table represents the number of paired ratings on a given attribute/competency. Rollups and averages are provided at the overall category level. Means (*M*) and standard deviations (*SD*) for all

ratings are reported (sample size for the reported descriptive statistics ranges from 131 to 135 ratees due to listwise deletion).

Building on these data, ICCs were calculated to evaluate interrater reliability for each LRM attribute/competency (cf., McGraw & Wong, 1996; Shrout & Fleiss, 1979). A one-way random effects model with multiple raters (average measures) was used due to the lack of consistent raters for all ratees. This type of model is based on absolute agreement among raters and typically yields the lowest (least favorable) coefficients compared to other ICC models. In an effort to retain a sufficient sample size for the analyses ($N > 30$), data only from the first three (of the five possible) raters were used to calculate ICC, resulting in a smaller dataset than reported for the percent agreement data shown above.

The resulting ICCs for ratings on each item are shown in Table 4. The ICCs reflect the degree of agreement among raters and indicate how similarly they rated individuals on a given attribute/competency. Guidelines for interpreting ICC values vary. For instance, on the more conservative side, ICC values of less than .40 indicate poor interrater reliability, values between .40 and .59 indicate fair reliability, values between .60 and .74 indicate good reliability, and values above .75 indicate excellent reliability (Cicchetti, 1994). In contrast, other guidelines note that values between .2 and .4 might be considered fair reliability (e.g., Altman, 1990). While these guidelines were used as a reference, such categorizations were not the focus of this research, especially in light of the fact that these data were acquired outside of the target context for which E-LARA was intended. Instead, the results helped to identify potentially problematic areas that may warrant further testing and eventual revision.

As reflected by the ICC values in Table 4, a greater degree of interrater reliability was found for the following attributes/competencies (ordered highest to lowest): Leads Others, Builds Trust, Fitness, Extend Influence, Leads by Example, Creates a Positive Environment, Communicates, and Prepares Self. The results suggest that these attributes/competencies were rated similarly across raters. A more moderate degree of reliability was found for Military and Professional Bearing, Stewards the Profession, Sound Judgment, Gets Results, Confidence, Innovation, Develops Others, Interpersonal Tact, Expertise, and Discipline. While the ICC values are within the acceptable range, the confidence intervals should be considered when interpreting the interrater reliability for these attributes/competencies. Weaker interrater reliability was found for ratings on the Army Values, Warrior Ethos, Mental Agility, Empathy, and lastly, Resilience, which had the lowest ICC value. Notably, of these five items, three were in the higher-level category of Character. Confirming these results, when rolled up to the higher-order conceptual categories, Character had the lowest value ($ICC_{mean} = .33$), whereas the other items were all over a value of .4, with the highest being for Leads ($ICC_{mean} = .67$) followed by Develops ($ICC_{mean} = .57$) and then Achieves ($ICC_{mean} = .53$).

Overall, these data suggest that there is promising evidence for agreement both as measured by percent agreement and ICC on most items. However, evidence for agreement was less in the category of Character, echoing the results found in the FTX use case noted above. In particular, while percent agreement values were all over 70% within one point, in terms of ICC, the Character sub-attributes (Army Values, Warrior Ethos, and Empathy) stood out as being more problematic, as well as Mental Agility (part of Intellect) and Resilience (part of Presence). It is possible that these areas are difficult to capture and observe consistently, in particular for new Soldiers who may not yet have been fully able to both experience and observe critical

challenges to items like Character, the need for mental agility, or the need for resilience. Yet, despite it being the case that the setting explored here does not fully match the ultimate intended use case (e.g., use of peers, limited LRM familiarity), moving forward these data indicate that further analysis and revision of these items may be warranted as additional data are collected.

Table 4

Interrater Reliability at the Attribute/Competency Level

Attribute/Competency	***ICC***	**95% Confidence Interval** Lower Bound, Upper Bound	***F***	***df1, df2***	***Sig.***
Character	***.33***				
Army Values	.39	-.01, .38	1.63	40, 82	.03
Empathy	.21	-.32, .55	1.26	41, 84	.18
Warrior Ethos	.32	-.15, .62	1.47	38, 78	.08
Discipline	.41	.03, .66	1.70	42, 86	.02
Presence	***.47***				
Military and Professional Bearing	.59	.33, .76	2.46	43, 88	.0001
Fitness	.69	.49, .82	3.20	45, 92	.0001
Confidence	.52	.21, .72	2.09	42, 86	.002
Resilience	.08	-.54, .48	1.09	40, 82	.37
Intellect	***.46***				
Mental Agility	.32	-.13, .61	1.48	41, 84	.07
Sound Judgment	.54	.24, .74	2.18	42, 86	.001
Innovation	.51	.15, .74	2.04	34, 70	.006
Interpersonal Tact	.46	.08, .70	1.86	36, 74	.01
Expertise	.46	.08, .70	1.85	38, 78	.01
Leads	***.67***				
Leads Others	.70	.52, .82	3.35	48, 98	.0001
Builds Trust	.70	.52, .82	3.34	48, 98	.0001
Extends Influence Beyond the Chain of Command	.69	.50, .82	3.23	47, 96	.0001
Leads by Example	.67	.46, .81	3.07	43, 88	.0001
Communicates	.60	.36, .76	2.52	48, 98	.0001
Develops	***.57***				
Creates a Positive Environment	.66	.45, .80	2.93	45, 92	.0001
Prepares Self	.60	.34, .77	2.52	42, 86	.0001
Develops Others	.47	.13, .70	1.89	42, 86	.006
Stewards the Profession	.55	.25, .74	2.22	42, 86	.001
Achieves	***.53***				
Gets Results	.53	.23, .72	2.10	45, 92	.001

Note. Mean ICCs were also calculated at the overall category level.

Discussion and Conclusion

This report describes the development and evaluation of a tool designed to assess and track Soldier performance on the attributes and competencies contained within the Army's LRM over time. E-LARA provides an efficient way for leaders to keep track of their Soldier's

performance over time. Specifically, E-LARA facilitates actionable on-the-spot feedback as well as specific descriptions of performance to be used in formal counseling sessions and evaluations (e.g., the NCOER). Throughout our data collection efforts, we remained adamant with regard to the protection of our Soldiers' data. We realize that it is of utmost importance to put into place certain precautions for doing so. Even with the closed systems we utilized, we directed all users and raters to use non-identifying methods such as nicknames, numeric digits and de-identified data (using unique alphanumeric identifiers) when we obtained data from other sources. This is perhaps the greatest challenge we, and anyone else developing, testing and implementing this type of technology, face in transitioning a product from demonstrational use to operational use. The Army needs to determine how these types of tools can be operationally used while safeguarding PII data.

As described, usability and utility data were collected to assess the degree to which E-LARA was perceived as an easy-to-use tool from which its users could also derive benefits. As demonstrated by the high SUS scores, Soldiers are able to easily navigate the application and determine how to use it to fit their needs. However, not all participants agreed that they would recommend this tool to others. Although based on a small sample size ($N = 5$), these data suggest that while the tool is usable, the Soldiers may not see the utility in it. In order to understand fully how E-LARA may be beneficial over time, it is imperative that data are collected over a longer amount of time. It may be that within the timeframe that data were collected, the Soldiers did not enter enough data for the tool to be useful to them. In addition, different events (e.g., completing NCOERs) may allow Soldiers to better see how and when E-LARA could help create efficiencies in their work. Finally, because the dashboard component was not running at the time the data were collected, it remains to be seen how that piece may influence the perceived usability and utility of the application.

The second type of evaluation data collected was interrater agreement and reliability data to demonstrate that the measures contained within E-LARA promote consistency in ratings. Overall, across two use cases (the FTX and the peer assessments), the results are favorable and demonstrate that raters generally agree on what constitutes good, average, and poor performance. Across both uses, patterns emerged demonstrating that certain attributes and competencies may be easier to rate than others. Specifically, when looking at both percent agreement and the ICCs, Character emerged as the most difficult to rate consistently. The inconsistent ratings of Character may have occurred for a few reasons. First, it may be that it was not possible to observe varying gradients of performance concerning Character in the settings used for this analysis. Alternatively, it may be that the scales created for Character need to be improved. By collecting additional data over time, this issue may be clarified. The data also point to the fact that not all attributes and competencies are salient in every setting. This finding again demonstrates that data must be collected over time, across multiple events, to be able to capture the full range of attributes and competencies for an individual.

Taken together, the results from the usability/utility analysis and the rater agreement and reliability analysis point to E-LARA as a promising tool for assessing and tracking Soldiers' performance over time with respect to the LRM. By using E-LARA and the accompanying dashboard consistently across various events, leaders should be able to form a clear picture of where an individual Soldiers stands in regard to leadership strengths and weaknesses. Ultimately, these data can be rolled up from the individual to unit level to provide leaders at

higher echelons with an informed picture of where additional development may be needed to facilitate mission accomplishment.

References

Altman, D. G. (1990). *Practical statistics for medical research.* London: Chapman & Hall/CRC press.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human Computer Interaction, 24*(6), 574–594.

Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies, 8*(2), 29-40.

Cicchetti D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290.

Dein, J. P., Ingurgio, V. J., Ratwani, K. L., Diedrich, F., & Flanagan, S. (2019). Tools and measures for NCO talent assessment. *NCO Journal,* July 2019. https://www.armyupress.army.mil/Journals/NCO-Journal/Archives/2019/July-2019/NCO-Talent-Management/

McGraw K. O, & Wong S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.

Shrout, P. E., & Fleiss J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Steiner, D. D., & Rain, J. S. (1989). Immediate and delayed primacy and recency effects in performance evaluation. *Journal of Applied Psychology, 74*(1), 136.

Tucker, J. S., Uhl, E. R., Allen, J. L., Kochert, J. F., Pitts, K. P., Brimstin, J. A., …Ledford, C. B. (2019). *Infantry One Station Unit Training Transformation: Phase I Findings* (ARI Research Report 2025). Arlington, VA: US Army Research Institute for the Behavioral and Social Sciences. (DTIC #AD1080564).

U.S. Department of the Army, Headquarters (2012). *Army leadership (ADP 6-22).* Washington, DC.

U.S. Department of the Army, Headquarters (2015). *Leader development (FM 6-22).* Washington, DC.

U.S. Department of the Army, Headquarters (2019a). *The Army people strategy*. Washington, DC.

U.S. Department of the Army, Headquarters (2019b). *Army leadership and the profession* (ADP 6-22). Washington, DC.

U.S. Department of the Army, Headquarters Army Training and Doctrine Command (2018). *The U.S. Army in multi-domain operations 2028 (TRADOC Pamphlet 525-3-1)*. Washington, DC.

APPENDIX A

LEADERSHIP ATTRIBUTE AND COMPETENCY RUBRIC

Complete Definitions and Rubric

LEADS	
Leads Others	Motivates, inspires and influences others to take initiative, go beyond self, work toward a common goal, and achieve objectives
Builds Trust	Establishes a positive environment that encourages commitment built on respect and shared experiences
Extends Influence Beyond the Chain of Command	Influences and coordinates with others beyond direct lines of authority to accomplish objectives
Leads by Example	Serves as a decisive leader and positive role model who upholds standards and Army Ethic (e.g., legal and moral Army principles)
Communicates	Clearly expresses ideas to effectively translate goals into actions; actively listens to others
DEVELOPS	
Creates a Positive Environment/Fosters Esprit de Corps	Establishes and maintains positive attitudes that support teamwork, fairness, open/honest communication, and personal responsibility
Develops Others	Encourages and supports others to grow as individuals and teams to meet future unit and Army needs
Prepares Self	Maintains and continues to improve the expertise and skills required for role, unit, and profession
Stewards the Profession	Makes decisions and takes action to improve the organization beyond current unit and position.
ACHIEVES	
Gets Results	Ensures results consistently meet expected mission end state and are accomplished in accordance with the Army Ethic (e.g., legal and moral Army principles)
CHARACTER	
Army Values	Abides by the Army Values in all decisions and actions
Empathy	Considers and responds to another's point of view and feelings and provides care
Warrior Ethos	Applies the spirit of the Ethos across all activities to succeed with honor
Discipline	Follows, demonstrates, and promotes sound practices in administrative, organizational, training, operational, and legal duties
PRESENCE	
Military and Professional Bearing	Projects a commanding presence and professional image of authority (e.g., outward appearance, attitude, actions, and words)
Confidence	Projects certainty in self and unit; calm and collected; possesses control of emotions

Fitness	Having sound health, strength, and endurance that support emotional health and effective thinking under prolonged stress
Resilience	Recovers quickly from adversity or injury while accomplishing mission
INTELLECT	
Mental Agility	Applies multiple perspectives and approaches; anticipates or adapts to ever-changing conditions
Sound Judgment	Makes sound and timely decisions; makes assessments according to strengths and weaknesses applicable to the situation and draws realistic conclusions
Interpersonal Tact	Possesses an awareness of how others view him/her; effectively interacts with others
Innovation	Introduces something new and creative as appropriate to situation; original in thoughts and ideas
Expertise	Possesses knowledge, facts, beliefs, and logical assumptions necessary for technical and tactical competence

LEADS	Developmental Need	Standard	Strength
Builds Trust	• Does not treat others with basic fairness and respect • Fails to address problems caused by team members that threaten trust in the unit • Makes no effort to build positive relationships with subordinates and teammates (e.g., does not get to know Soldiers) • Delegates responsibility without oversight	• Treats others with fairness and respect • Addresses problems as they arise but before they can cause trust issues in the unit • Uses shared experiences to relate to others and build positive relationships • Appropriately delegates tasks with limited or slow follow-up	• Addresses potential problems before they arise and cause trust issues in the unit • Retains responsibility and verifies that delegated tasking meets mission objectives by engaging in a timely follow-up even in unfamiliar conditions
Communicates	• Information does not get passed to everyone • Conveys information in a manner that is not organized, clear, accurate, or timely • Does not listen and/or interrupts others	• Recognizes and resolves misunderstandings • Informs higher and lower headquarters, superiors and subordinates as required • Conveys complete and accurate information but may not be concise • Listens, but may not clarify or take notes	• Shares complete and accurate information in a timely, concise manner, and ensures updates are provided as needed • Informs and verifies understanding with higher and lower headquarters, superiors and subordinates • Engages in active listening (e.g., clarifies; takes notes) • Proactively gets input from the team

LEADS	Developmental Need	Standard	Strength
Leads Others	• Fails to delegate, or delegates but loses control of subordinates, resulting in task/mission failure • Fails to influence others to accomplish objective • Fails to ensure that self and subordinates meet standards • Unaware or dismissive of effects of mission fulfillment on subordinates' mental, physical, and emotional attributes	• Delegates appropriately but may not always follow up after delegating tasking and/or may sometimes micromanage when not appropriate/needed • Influences others to accomplish the immediate task but does not utilize the most effective methods • Ensures self and subordinates meet standards • Monitors effects of mission fulfillment on mental, physical, and emotional attributes of subordinates; expresses understanding	• Retains responsibility and verifies that delegated tasking meets mission objectives by engaging in a timely follow-up without unnecessarily micromanaging • Matches techniques of influence (e.g., personal appeals, collaboration, inspiration) to situation and individual needs • Pushes self and subordinates to exceed standards; takes responsibility for poor performance and addresses it appropriately • Monitors effects of mission fulfillment on mental, physical, and emotional attributes of subordinates, and takes appropriate action for relief as needed
Extends Influence Beyond the Chain of Command	• Unable to establish common ground and resolve conflict • Ineffectively leverages resources to accomplish tasks, missions • Works only within immediate chain of command for task accomplishment	• Maintains unity by establishing common ground and helping resolve conflict • Only leverages local resources; reaches beyond only when there is a problem • Works with peers in immediate adjacent units to facilitate task accomplishment	• Proactively builds unity by establishing common ground and resolving potential conflict • Proactively builds working relationships to effectively accomplish task, mission by leveraging full range of available resources • Works with appropriate individuals across units to facilitate task accomplishment

LEADS	Developmental Need	Standard	Strength
Leads by Example	• Fails to display high standards; violates the Army Ethic • Ignores the perspective and ideas of others • Does not engage in honest or appropriate communication	• Models high standards and the Army Ethic under standard conditions; completes training to standard • Considers new ideas and perspectives but does not fully leverage them to enhance unit performance • Models honest and appropriate communication	• Models high standards and the Army Ethic even under difficult circumstances; does the right thing even when thinking no one is watching • Seeks diverse and new ideas and leverages them to enhance unit performance • Encourages and models honest and appropriate communication

DEVELOPS	Developmental Need	Standard	Strength
Creates a Positive Environment/Fosters Esprit de Corps	• Does not ensure subordinates have the required time and resources to do their job • Frequently complains; gossips • Lacks consistency in attitude and behavior • Does not listen to others' issues and suggestions • Is non-inclusive; may bully others • Unfairly administers discipline and/or rewards	• Ensures subordinates have the time and resources to do their job • Consistent in attitude and behavior • Listens to others' issues and suggestions and takes appropriate action • Inclusive and does not bully others • Fairly administers discipline and/or rewards	• Considers costs and benefits while ensuring subordinates have the time and resources to do their jobs • Consistently sets and maintains a positive and inclusive climate even when under pressure • Keeps complaints/gossip to a minimum in a professional way • Listens to others issues and suggestions and takes action to resolve problems; proactively develops others; takes ownership of subordinates' development

DEVELOPS	Developmental Need	Standard	Strength
Develops Others	• Solves problems for subordinates • Provides counseling to individuals and/or teams but does not address improvements (e.g., 'great job' and sugarcoats); provides feedback that is not actionable • Attempts to let subordinates work through problem but lacks patience to allow subordinates to fully solve it (jumps in prematurely)	• Provides counseling to individuals and/or teams with a balance of good and bad feedback but does not provide actionable feedback on how to improve • Coaches and has sufficient patience for subordinates and/or teams to solve problems • Waits until there is an apparent problem to provide mentoring and coaching; misses some opportunities to develop subordinates	• Provides counseling that is balanced and results in an actionable plan focused on achieving current and future performance objectives • Patiently coaches others (peers and subordinates) and provides opportunities for Soldiers and/or teams to succeed • Anticipates and address developmental problems before they occur in Soldiers; proactively sets conditions that foster growth and promotability
Prepares Self	• Does not participate in self-development courses • Lacks self-awareness about own weaknesses • Consistently unprepared to execute mission, training events	• On schedule, making progress toward career goals (e.g., promotion board) • Recognizes weaknesses and takes the time to improve by working on weaknesses during own time • Always prepared to execute mission, training events	• Ready or ahead of schedule to go to promotion board • Proactively seeks out experiences for advanced training (e.g., Gainey Cup; ARC) • Anticipates and plans for future mission and position requirements
Stewards the Profession	• Does not fully identify opportunities for unit or subordinate development • Does not put a lot of thought into what happens beyond local training exercises • Helps his/her own immediate team/unit accomplish a task	• Helps higher-level unit succeed (e.g., is a team player) • Identifies opportunities for unit or subordinate development (e.g., key assignments, PME attendance) • Demonstrates good planning and	• Helps higher-level unit succeed by identifying areas of opportunity; shares process improvements to benefit future units; does not hold information just for him/herself or immediate team • Understands big picture and engages in actions for the greater good

DEVELOPS	Developmental Need	Standard	Strength
	• Unnecessarily wastes time and/or resources; does not prioritize	forethought in how organizational resources are to be used	• Optimizes time and resources for both the short and long term benefit of the unit

ACHIEVES	Developmental Need	Standard	Strength
Gets Results	• Routinely fails to meet end state within commander's intent and/or violates the Army Ethic while doing so • Does not fully leverage strengths of the team	• Meets end state within commander's intent and consistent with the Army Ethic but may not be efficient or timely • Leverages strengths of the team but may not fully account for long term effects (e.g., unnecessarily smokes team members)	• Always meets end state within commander's intent in a timely manner and consistent with the Army Ethic • Leveraging strengths of the team and efficiently using resources, while considering long term effects as appropriate • Routinely uses additional time available to proactively prepare for the next action

CHARACTER	Developmental Need	Standard	Strength
Army Values	• Fails to support leadership and/or lets teammates fail when in leadership roles • Fails to meet obligations, accomplish tasks, or fulfill responsibilities • Is intolerant toward diversity; does not give others a chance; creates a counterproductive environment • Does not help others or seeks recognition/personal gain for doing so • Makes immoral or unethical decisions • Is unable to overcome challenges and physical fears; Does not stand firm on values or principles	• Is a team player who supports leadership • Meets obligations and responsibilities individually and as member of a team • Remains open to different perspectives; listens to others; maintains positive and inclusive climate • Helps others and does not expect recognition for doing the job right • Consistently makes decisions that are morally and ethically sound • Works through challenges or physical fears to accomplish mission; Stands firm on values or principles regardless of circumstances	N/A - Followed Go/No-Go approach
Empathy	• Bullies or excludes others • Does not listen to others' perspectives • Fails to account for subordinates' strengths and weaknesses; uses a one-size-fits-all approach • Fails to act when subordinates/peers are struggling despite opportunities to intervene	• Is inclusive/supportive, without compromising task/mission requirements • Actively listens to others' perspectives (e.g., demonstrates understanding; asks clarifying questions, provides comments or words of support) • Considers subordinates' strengths and weaknesses when planning tasks or delegating • Helps peers/subordinates when they are struggling	N/A - Followed Go/No-Go approach
Warrior Ethos	• Is easily discouraged; quits or gives up • Places personal needs above mission • Does not strive to improve him or herself or team/unit based on failures • Does not help others	• Does not quit under challenging conditions • Places mission above personal needs • Attempts to learn from negative events • Helps others even under adverse conditions	N/A - Followed Go/No-Go approach

CHARACTER	Developmental Need	Standard	Strength
Discipline	• Impulsive (e.g., lacks personal control) • Takes the easy wrong over the hard right • Does not always consider legal, moral, and ethical consequences • Fails to consistently meet or hold others accountable to standard (e.g., legal, moral, physical)	• Maintains composure even when under stress • Does what is right; lives the Army values • Considers legal, moral, and ethical consequences • Holds self and others accountable to meet or exceed standard (e.g., legal, moral, physical)	N/A - Followed Go/No-Go approach

PRESENCE	Developmental Need	Standard	Strength
Military and Professional Bearing	• Fails to have uniform squared away; has poor hygiene • Looks at ground when speaking; has a shaky voice • Does not communicate or act in a courteous manner	• "Looks the part" of a Soldier (e.g., cleanly shaven; clean haircut, appropriate uniform) • Follows customs and courtesies; adheres to Army standards • Communicates and acts calmly and effectively	• Models appropriate customs and courtesies even when no one is looking • Communicates and acts calmly and effectively while motivating/energizing others • Encourages peers and subordinates to look the part of a Soldier
Confidence	• Unable to maintain composure/self-control • Unable to a make decision or rushes to incorrect decision • Rejects (e.g., argues) constructive criticism when given • Continuously seeks input without deciding, acting (e.g., lack competence) • Overly confident; manages all aspects of planning, execution without accounting for subordinate suggestions	• Maintains composure/self-control under standard conditions, but struggles as stress and ambiguity is introduced • Makes sound decisions under standard conditions, but may not be timely or may be likely to waver when pressed • Accepts constructive criticism but does not/is slow to change • Task, purpose and end state is overly vague or	• Maintains composure/self-control while under stress or with little information • Makes timely and sound decisions while solving complex problems • Incorporates constructive criticism • Maintains command and control even under difficult conditions and effectively conveys task, purpose, and intent

PRESENCE	Developmental Need	Standard	Strength
		too complicated, suggesting uncertainty	
Fitness	• Does not meet minimum physical requirements • Lets performance suffer under stress (e.g., gives up easily) • Does not follow adequate PT plan	• Meets minimum physical requirements • Occasionally exhibits difficulty performing under pressure • Follows adequate PT plan	• Exceeds APFT standard • Consistently endures and performs to a high standard under prolonged stress • Helps others develop adequate PT plan
Resilience	• Is unable to bounce back, shuts down, and/or loses organizational/mission focus after a setback or negative event • Unit fails to bounce back after setbacks/negative events	• Recovers from setbacks and negative events, but is slow to apply and demonstrate improvement • Maintains organizational/mission focus • Unit bounces back after setbacks/negative events but relies solely on the leader	• Quickly recovers and learns from setbacks or negative events to improve performance • Maintains organizational/mission focus even after extreme setbacks • Unit bounces back after setbacks/negative events and fully leverages capacity of all members

INTELLECT	Developmental Need	Standard	Strength
Mental Agility	• Rigidly follows initial plan causing mission failure; fights the plan, not the battle • Constantly surprised by unexpected conditions; lacks ability to think ahead; does not plan for contingencies	• Tries to develop solutions individually or only involves a limited number of perspectives, resulting in few contingencies • Identifies main problem and implements solutions, but slowly and/or may not be optimal • Anticipates unexpected events; solves local problem	• Identifies problems and changes behavior in an optimal and timely manner in response to ambiguous, complex or changing conditions • Stays one step ahead of problem, identifies second and third order effects, and takes opportunities as they emerge • Collaboratively develops solutions with multiple perspectives and contingencies, leading to optimal execution
Sound Judgment	• Fails to assess the situation; does not make timely or sound decisions • Does not ask clarification questions or seek more information if unsure how to proceed • Cannot explain the "why"	• Independently draws realistic conclusions but may not assess all relevant pieces of information (e.g., METT-TC); focuses mainly on the tactical evidence • Explains the “why” • Makes timely decisions	• Effectively seeks and integrates multiple relevant pieces of information (e.g., METT-TC) to make an informed decision before taking action; considers consequences of decision • Justifies decision making based on a sound assessment of the situation • Takes prudent risks when appropriate; uses time wisely and prioritizes effectively, even under stress or time pressure
Interpersonal Tact	• Is unable to adjust tone and interaction style • Loses self-control • Does not pay attention to non-verbal signals from others (e.g., eye rolling); does not listen to others (no turn taking) • Is intolerant toward diversity	• Maintains self-control under standard conditions • Makes attempts to adjust tone based on needs and perceptions of others but does not always do so effectively • Reacts to non-verbal/social cues appropriately • Accepts diversity but	• Effectively adapts interaction style across multiple contexts • Embraces and leverages diversity to enhance unit performance/mission • Maintains self-control under stress and adversity

INTELLECT	Developmental Need	Standard	Strength
		is unable to fully leverage it	
Innovation	• Relies on traditional methods that may not work when faced with challenging circumstances • Relies solely on the creativity of others to solve problems	• Attempts to adjust and try novel approaches but may not be fully effective, feasible, or needed • Independently develops new ideas without fully leveraging others	• Thinks past standard solutions to recognize feasible opportunities for improving situation/ process/performance; changes behavior and proposes new outcomes • Develops new ideas while building on others' ideas • Encourages and questions others' ideas to foster new perspectives as appropriate
Expertise	• Does not know how to or cannot apply required tactical/technical procedures	• Displays expected level of tactical/technical expertise consistent with role/position and event • Recognizes own level of expertise and takes appropriate action to learn • Trains subordinates to standard	• Employs individuals, teams, organizations effectively to fulfill mission objectives • Seeks ways to expand knowledge and shares it with peers/subordinates • Trains subordinates to standard and also helps peers as necessary

APPENDIX B

DATA COLLECTION MEASURE

Directions: Place a check mark in the box that best represents your rating of your experience using the mobile application).

	Strongly Disagree				Strongly Agree
1. I think that I would like to use the app frequently.	1	2	3	4	5
2. I found the app to be unnecessarily complex.	1	2	3	4	5
3. I thought the app was easy to use.	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use the app.	1	2	3	4	5
5. I found the various functions in the app to be well integrated.	1	2	3	4	5
6. I thought there was too much inconsistency in the app.	1	2	3	4	5
7. I would imagine that most people would learn to use the app very quickly.	1	2	3	4	5
8. I found the app to be very cumbersome to use.	1	2	3	4	5
9. I felt confident using the app.	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with this app.	1	2	3	4	5

Free response

1. If I could change or remove a feature in the app, it would be …

2. If I could add anything to the app, it would be …

3. Would you recommend the current version of the app to members of your unit, colleagues, etc.?