Cognitive Skill Assessment in a Virtual Environment

Allison K. Hancock, Jennifer K. Phillips
Cognitive Performance Group
Independence, OH
allison@cognitiveperformancegroup.com
jenni@cognitiveperformancegroup.com

Natalie B. Steinhauser[1]
Naval Air Warfare Center
Training Systems Division
Orlando, FL
natalie.steinhauser@navy.mil

James Niehaus
Charles River Analytics
Cambridge, MA
jniehaus@cra.com

ABSTRACT

Adaptability is a critical skill necessary for current and future operating forces. In complex arenas where problems are shifting continually, adaptability is essential for achieving mission success. However, efforts to operationalize and assess the construct of adaptability have fallen short. Existing adaptability assessments do not meet the military's need for an objective, tactically realistic, and diagnostic capability. In addition, many are resource intensive to maintain and administer, and therefore are not sustainable. The Assessment Toolkit for Leader Adaptability Skills (ATLAS) addresses these challenges and extends the research and development in the area of complex skill assessment.

The ATLAS tool measures adaptability in tactical vignettes representing mission sets faced by maneuver small unit leaders. To eliminate the need for a resource-intensive human rater scoring process, five standard and objective measures of adaptability are captured from user actions and scored using automated algorithms. The measures of adaptability align with steps of a framework hypothesizing the flow of cognitive processes that must occur for a favorable adaptive outcome and are the basis for the diagnostic feedback provided to users.

To date, the measures and prototype vignettes have been instantiated in a Unity-based virtual simulation environment for hypothesis testing. Pilot testing resulted in advancements in the development of detailed, diagnostic feedback that will support the transition from scoring output to targeted training recommendations. Future phases of research will validate the adaptability measures in a large-scale field test and examine the validity of the measures when a rapid content generation function is applied. This function will automatically modify the repository of vignettes for increased sustainability of the tool. The outcomes of this effort contribute to the research community by providing a framework, method, and automated measures for assessing individual developmental needs for adaptability in the context of engaging and realistic simulations.

ABOUT THE AUTHORS

Dr. Allison K. Hancock is a Senior Scientist at the Cognitive Performance Group; specializing in operational military research and human performance assessment. Dr. Hancock has a Ph.D. in Educational Psychology with specialization in Sport and Exercise Psychology as well as Measurement and Statistics and a M.A. in Applied Experimental Psychology. She has over ten years of experience in leading research projects and managing multiple contracts with distributed contract teams for the Department of Defense in the areas of human performance, rapid training development, cognitive skills training, program evaluation, and assessment.

Jennifer K. Phillips is a Senior Scientist at the Cognitive Performance Group. Her research interests include skill acquisition, adaptive training, cognitive performance improvement, and the nature of expertise. Ms. Phillips applies cognitive task analysis and related techniques to model performance across the levels of proficiency, designs learning solutions including decision-centered training scenarios and facilitation techniques, and develops metrics for cognition and decision making. She is currently supporting Marine Corps initiatives in the areas of small unit decision making, instructor professional development, and modernization of the Intelligence, Surveillance, and Reconnaissance enterprise.

Natalie B. Steinhauser is a Senior Research Psychologist at NAWCTSD in the Basic and Applied Training and Technology for Learning and Evaluation (BATTLE) laboratory. She has 14 years of Navy experience conducting

[1]The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.

training effectiveness evaluations, developing adaptive training and intelligent tutoring systems, and conducting decision-making research. She holds a M.S. in Modeling and Simulation from the University of Central Florida, and a B.S. in Human Factors/Applied Psychology from Embry-Riddle Aeronautical University.

Dr. James Niehaus is a Principal Scientist at Charles River Analytics. His areas of expertise include artificial intelligence and training systems, and he is currently the principal investigator on multiple efforts developing new training technologies. He leads projects to: develop psychological resilience applications for the Marines, enhance physical therapy with video games and virtual coaches, develop a game-based training system for missile defense operators, and investigate the neural, cognitive, and behavioral relationship between implicit learning and intuition. Dr. Niehaus has a B.S. in computer science from the College of Charleston and a Ph.D. in computer science from North Carolina State University.

[1]The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.

Cognitive Skill Assessment in a Virtual Environment

Allison K. Hancock, Jennifer K. Phillips
Cognitive Performance Group
Independence, OH
allison@cognitiveperformancegroup.com
jenni@cognitiveperformancegroup.com

Natalie B. Steinhauser[1]
Naval Air Warfare Center
Training Systems Division
Orlando, FL
natalie.steinhauser@navy.mil

James Niehaus
Charles River Analytics
Cambridge, MA
jniehaus@cra.com

INTRODUCTION

Adaptability is a complex cognitive skill necessary for current and future operating forces to successfully achieve their missions in multifaceted environments where problems are shifting continuously. Previous work on small unit leader decision making identified adaptability as one of nine key performance areas of highly proficient small unit leaders (Ross, Phillips, & Rivera, 2013). A key performance area is a critical task, behavior, or attribute related to job performance. The Marine Corps' School of Infantry identified adaptability as one of two primary performance goals for infantry small unit leaders. Yet despite its criticality to the tactical success of the Marine squad leader, no validated, objective adaptability assessment capability exists that meets the needs of the military.

A review of the research literature and military technical reports has not yielded adaptability measures appropriate for Marine Corps needs. Perspectives on how to approach the assessment of adaptability fall into two categories: *domain-general* or *domain-specific* (Kozlowski & Rench, 2009). A *domain-general* approach postulates that adaptation is a set of traits and/or skills that can be applied in any area of performance, and it does not recognize the unique problems or developmental path in a specific task or job context. The *domain-specific* approach views adaptation as a process and concentrates on the knowledge and skills in a specific domain of practice (Baard, Rench, & Kozlowski, 2014). The domain-specific approach is more appropriate for capturing the adaptive performance of a Marine small unit leader. For a Marine squad leader, adaptability can be defined as "The ability to fluidly apply knowledge and tactical principles across situations, or alter one's plans, actions, or decisions when the situation, environment, or circumstance has changed, while still accomplishing the mission or intent" (Ross et al., 2013, p 7). This description of the adaptive performance required by the squad leader highlights the requirement for domain-specific knowledge and skills in order to be effective.

The *domain-specific* approach to assessment can yield more direct cognitive and behavioral indicators of performance. However, the few existing assessment examples are resource intensive to develop, have required natural language as user responses, and are limited by scenario-specific metrics that do not generalize to new scenarios.
Approaches using single skill assessment, such as Think Like A Commander (TLAC) critical indicators or the Situational Awareness Global Assessment Technique (SAGAT; Endsley, 2000) situation awareness assessment are helpful in understanding critical contributors to adaptive expertise, but are resource intensive and do not in themselves constitute adaptability assessment. One Situational Judgment Test (SJT) designed to measure adaptability, the Adaptive FORCE Program produced for a Marine Corps Training Command school (OSD P&R, 2009), was not predictive of tactical decision-making performance or time in service in previous research (Ross, Phillips, & Knarr, 2015).

The U.S. Army's Asymmetric Warfare Group (AWG) produced a scenario-based adaptability test as part of the Soldier Performance and Effective, Authentic Response Study (SPEARS; AWG 2016), which may be the most appropriate match to Marine squad leaders of the existing adaptability assessments. However, the SPEARS measure, like others of its kind, depends on subject-matter experts to rate Soldier performance along a scoring rubric. The use of human raters in an assessment can provide valuable feedback to the performer in addition to a performance score, however, the approach is labor-intensive in its requirement to train raters, routinely assess inter-rater reliability, and utilize rater time during participant task performance.

The sustainability of an assessment product depends on its continued relevance to the force, yet scenarios generated to provide the context for domain-specific assessments become obsolete, and are costly to develop, especially when they are instantiated into a virtual simulation environment. A large library of scenarios would be required for an

[1]The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.

assessment to be applied at regular intervals in order to continuously assess improvement over time. When missions and operational contexts of interest change, such as the shift from counterinsurgency operations in Iraq and Afghanistan to heightened preparation for near-peer threats, new scenarios are required to keep pace with the focus of the services. This reality argues for a rapid content generation capability to enable adaptability testing in new mission and terrain environments, as well as measures that are not scenario-specific and apply more generally across decision types and contexts.

The Assessment Toolkit for Leader Adaptability Skills (ATLAS) seeks to leverage virtual simulation, recent advances in automated decision metrics, and a procedural content generator to overcome the challenges associated with previous adaptability assessments. To summarize, the challenges include: (1) operationalizing the construct of adaptability in a manner such that it can be measured discretely; (2) providing diagnostic assessment outcomes that support training actions; (3) minimizing the costly need for human raters; and (4) producing a sustainable tool with a repository of scenarios for repeated testing. ATLAS assesses Marine small unit leader adaptability in the context of tactical vignettes that portray mission sets faced by maneuver small unit leaders. Using a theoretical framework of the cognitive processes in support of adaptability, ATLAS incorporates five automated measures of cognitive processes instantiated in a Unity-based tactical simulation environment. The use of a virtual environment for assessment affords the ability to capture cognitive performance data more efficiently than in the real world and to do so repeatedly to evaluate changes in performance over time. The incorporation of automated algorithms to score user responses relieves the need for a human rater and the potential for rater biases and errors.

The purpose of this paper is to describe the design and development of the ATLAS virtual simulation as a model for applying automation to the problem of measuring complex cognitive skills without forfeiting the richness of the context in which they are brought to bear. In the sections to follow, we describe how the construct of adaptability is operationalized into its hypothesized component steps to provide a standardized set of measures across contexts. We then discuss the design of the virtual simulation environment, including how users interact with the system and how the content is designed, to allow for a standardized and objective test as well as automated modification of the scenarios for sustainability. Finally, we conclude with a discussion of lessons learned on standardizing assessment of cognitive skills in a virtual environment and next steps to validating our approach.

OPERATIONALIZING ADAPTABILITY

The goal of this research and development is to measure adaptability in a manner such that assessment outcomes can be used to determine the additional training or remediation necessary to improve performance. If we focus only on the situation-specific adaptive behaviors, such as a change in tactics or the employment of assets, then the assessment is not effective in diagnosing the elements of cognition to be strengthened to improve future performance. With a diagnostic assessment of development needs, learning approaches can be designed for implementation in schoolhouse settings and individual shortcomings can be addressed with targeted instruction. With that end in mind, a previous phase of this research (Ross & Phillips, 2017) set out to identify how other researchers have addressed operationalization and measurement of this complex skill. In a review of the scientific literature, little consistency was found within either military or non-military programs of research with regard to the definition or measurement approach for the construct of adaptability. Using the literature review and outcomes from a study of performance requirements for Marine squad leaders, the initial phase of research identified an operational definition of adaptability for military small unit leaders and hypothesized a cognitive framework for adaptability incorporating discrete and measurable steps of the cognitive process.

[1]The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.

Adaptability is viewed as a competency that can be demonstrated behaviorally by the actions taken when circumstances change in a manner that (1) requires or (2) allows a change in plans to meet mission goals. It is defined as *fluidly modifying or changing one's planned actions when the situation has changed from what was expected, or when the typical approach or plan is rendered less effective than necessary* (Ross & Phillips, 2017). Three underlying cognitive processes are believed to support adaptive performance: change detection, anomaly detection, and cognitive flexibility. The model of the cognitive framework for adaptability, shown in Figure 1, defined the constructs of interest as follows:

- **Change Detection**: Attending to relevant aspects of the environment in order to perceive a difference in one or more elements in the situation, and interpreting that difference to support one's situational awareness, understanding of baseline, or immediate threat assessment.
- **Anomaly Detection**: Realizing through perceptual-cognitive processes that the presence or absence of elements or patterns of elements in the environment is off the baseline for that context or activity, and therefore requires more explicit reasoning to locate the source of the anomaly and understand its implications.
- **Cognitive Flexibility**: Applying knowledge and principles of tactics and leadership differently based upon the unique demands of the situation. The recognition that knowledge and schema learned in one context is relevant to other contexts and the flexible application of that knowledge and schema.

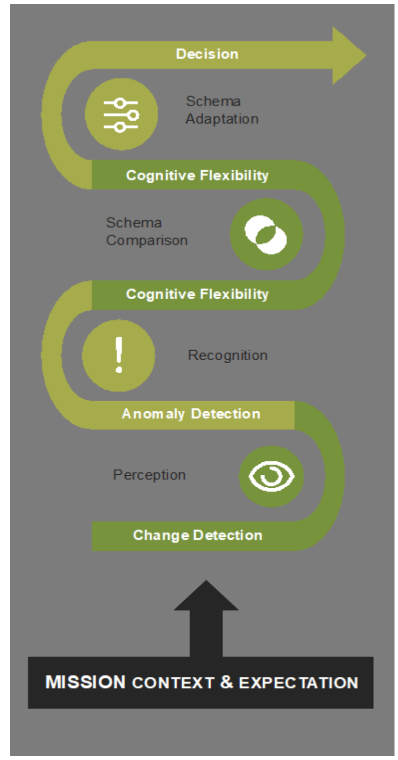

Figure 1. Hypothesized cognitive framework for adaptability.

The measurement objective, then, is to assess performance along each of the discrete steps hypothesized to comprise the cognitive framework for adaptability. Five measures were identified that align with each step in the cognitive process, are assessable using automated algorithms, and can be applied in a scenario-agnostic manner. Figure 2 illustrates the measurement framework applied in the virtual simulation. The steps of the adaptability framework relate to John Boyd's OODA—Observe, Orient, Decide, and Act—loop (1987) describing how individuals react to stimuli.

The first measure is *speed of reaction* to a trigger event. A trigger event is an occurrence that changes the nature of the situation in a way that impacts the mission. The speed of reaction measure is defined as the time from the trigger event to the user's first action demonstrating recognition that an adaptation is needed. The speed of reaction measure relates to the Observe step of OODA. This measure addresses the constructs of change and anomaly detection, by capturing the perception that a change has occurred and recognition that it has an impact on mission accomplishment.

The second measure is the *duration of adaptive activity*. This measure is defined as the total time between the first adaptive action after the trigger event has occurred and end of last adaptive action. This duration measure is hypothesized to provide an indication of the user's ability to compare schemas and assemble new schemas, or in other words, to retrieve relevant experience to assess the new situation and to re-plan. Individuals who possess a higher degree of cognitive flexibility and more schemas for comparison are expected to show lower duration of adaptive activity because they will be more decisive in their adaptations.

The third measure is the *quality of new situation assessment (SA)*. It is defined as participant prioritization of goals or concerns as compared to an expert referent and can be related to the Orient step of the OODA loop. This measure is an indication of one's knowledge and experience to produce an accurate new situation assessment given the perception of situational cues and detection of a change or anomaly.

The fourth measure is the *quality of adaptive actions*, which corresponds to the Decide step of the OODA loop in that it captures the participant's decisions about which adjustments to make in response to the changed situation. In the simulation, measuring the quality of adaptive action relies on a combination of automated assessments of actions against tactical rules and comparisons to expert referent where tactical rules cannot be represented in an algorithm.

[1]The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.

This measure indicates the user's ability to combine knowledge and experience to re-plan, which is another element of cognitive flexibility.

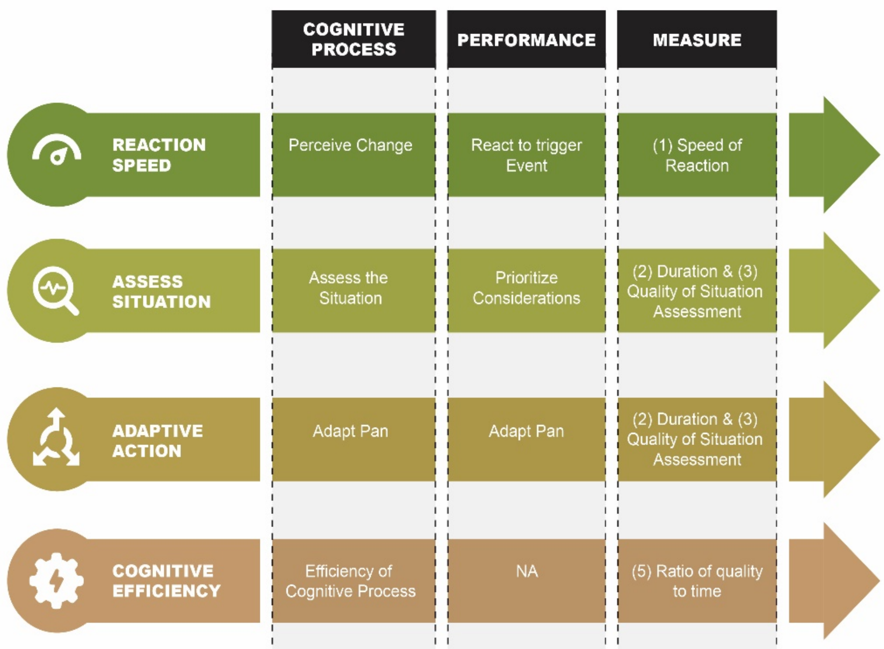

Figure 2. Measures of adaptability embedded in the ATLAS system.

The fifth and final measure hypothesized to differentiate adaptive performance is the *effectiveness of the adaptive process*. This measure, defined as the ratio of quality of activity (Measures 3 and 4) to duration of activity (Measure 2), does not correspond to a step in the cognitive framework but is viewed as a means of discriminating levels of proficiency. Users who make rapid decisions because they have a small repertoire of scripts and schema from which to choose should be differentiated from those who make rapid decisions because they intuitively recognize an effective course of action.

VIRTUAL SIMULATION DESIGN

The approach to capturing the five measures of adaptability in a virtual environment relies on three integrated elements of simulation design: (1) a set of representative vignettes representing an assortment of mission types and contexts, to ensure we are measuring the adaptive performance agnostic of situation type rather than an individual's degree of experience in the scenario type presented; (2) a standardized item structure, to maintain the integrity of the system as a repeatable test presenting uniform test questions to all participants; and (3) a standardized user interface producing response data suitable for automated algorithms and comparisons across participants. The challenges and lessons learned associated with this approach to adaptability assessment are described in this section.

Vignette Format and Context

The vignettes in the system are designed to provide a Marine squad leader mission and environmental context in which a change to the plan must occur. One challenge is to create vignettes that are realistic and immersive while maintaining a format and user interface that captures data to calculate measures of adaptability. Therefore, the graphics and interface of the system were designed to fully engage users in the simulation while controlling the amount of interactivity so that measurement can occur uniformly across participants. The vignettes are "on rails," meaning users interact with the simulation as if it is a movie playing out for them. In other words, the user does not influence the

[1]The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.

scenario as it plays out in a fully interactive manner, otherwise every participant would encounter a different situation to which to adapt. In addition, vignette design was constrained to a set of adaptive decisions typical for small unit leaders, but not representative of every adaptation possible.

The first step in vignette development was to identify a standard set of decision types that would be relevant across mission types and suitable for a robust assessment using automated algorithms. We gravitated toward decisions in which terrain reasoning skills would be applied. Recent advances in the Office of Naval Research (ONR) Human Performance, Training & Education (HPT&E) program have produced simulation-based algorithms measuring zones of concealment and lines of sight given positions on various terrains. These algorithms are currently used to assess route planning decisions based upon whether the route is concealed from enemy view, but may be extended to apply to decisions about where to place units and how to assign sectors of fire given friendly positions and enemy fields of observation as well as fields of fire. With those opportunities in mind, we identified the types of tactical decisions suitable for automated algorithms:

- <u>Line of sight</u>: cover, concealment, and camouflage
- <u>Traversability of terrain</u>: time it takes to move, effort to move, difficulty to navigate
- <u>Positioning of assets:</u> fields of fire, concealment, weapon system ranges and specifications, friendly unit locations, enemy unit locations
- <u>Route selection</u>: line of sight/concealment, traversability, ability to react to/engage enemy
- <u>Formations:</u> time of day, terrain type
- <u>Rates of movement:</u> time of day, concealment, terrain type/traversability
- <u>Communication protocol</u>: to whom, frequency, timing, what is reported, push versus pull
- <u>Call for Fire protocol</u>: position of target, timing, friendly positions

From the tactical decision types identified as suitable for measurement, three decision categories were selected for the simulation. These decision categories are expected to be applicable to every vignette:

1. Exchange Information. Decisions include communicating *to* a particular individual or requesting a communication *from* an individual. The participant, in the role of squad leader, may decide to report to the platoon commander or request guidance from the platoon commander; give information to a fire team leader or request information from a team leader; and so forth.
2. Provide Leadership. Decisions in this category are adjustments to how the squad is maneuvered and include adjusting the route, re-positioning fire teams, assigning fire team roles, selecting unit formation, setting battle positions, setting sectors of fire, and requesting a time-related change.
3. Request Support. Support request decisions include requesting specialized teams such as sensitive site exploitation or explosive ordnance disposal, or requesting indirect fire support by setting requested targets or calling for fire on a pre-planned target.

Once a standard set of decision types was selected for incorporation, the next step was to identify vignette mission contexts. One goal was to represent a variety of contexts, both in terms of the operation being conducted as well as the physical environment. It is critical that the system assesses adaptability across multiple mission types so the measurement is not confounded by participants' real-world experience in a specific context. For example, if we measure adaptability only in an Afghanistan-like counterinsurgency patrol mission, individuals who have deployed to Afghanistan are likely to perform better than those who have not. To that end, we selected vignettes representing combat patrols in rural, semi-wooded terrain, checkpoint operations in sparse desert vegetation, and offensive raids on a compound. These three vignette contexts were chosen from a repository of existing tactical scenarios to control costs associated with creating a new scenario from scratch and to cover a variety of contexts. Once the assessment system is fielded, users are expected to complete assessments within each of the three vignette contexts in a single test-taking session.

Next, the vignette format was developed. At the beginning of each vignette, the participant is given a fragmentary order and a short mission brief. Then, four vignette phases are presented as shown in the top row of Figure 3. Each phase contains distractor information and three levels of the trigger event. These triggers are progressively more obvious to the user, culminating in a trigger so apparent that the user is forced to make an adaptive response. The rationale for forced adaptation is to ensure every user provides response data for assessment points.

[1]The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.

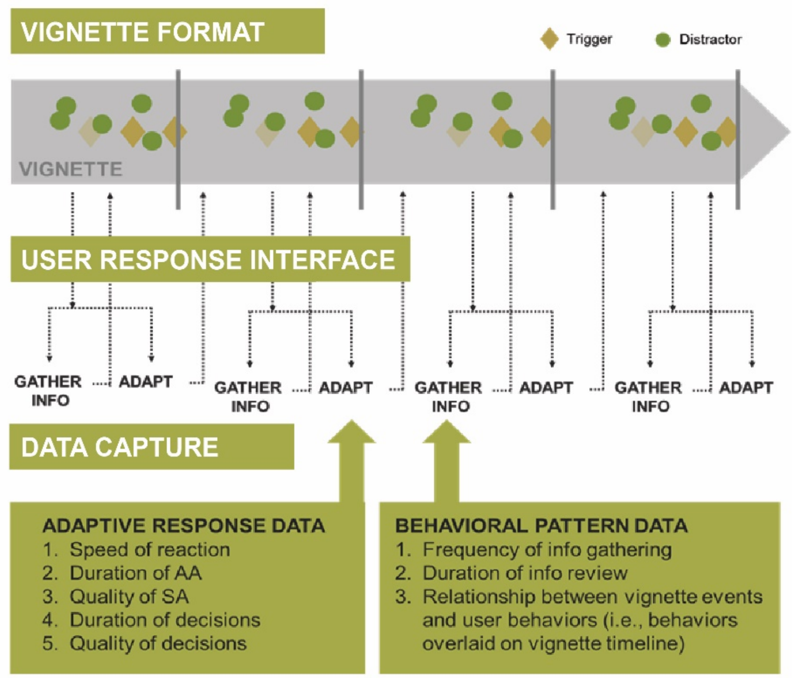

Figure 3. Integration of vignette format, user response interface and data capture.

Standard User Response Interface

As the vignette plays out, users interact using two broad response options—*Gather Information* and *Adapt*. The two broad response options contain subcategories of actions which branch out as appropriate for that action. For example, under *Adapt* is the *Exchange Information* set of decision options, and within that subcategory, the user must select with whom they would like to exchange information and then select what type of information will be exchanged.

Gather Information. The purpose of this broad response option is to give users access to information for reference or the ability to pull new information, as they would have under natural circumstances. Note that gathering information does not constitute a response in service of the five adaptability measures. Instead, it provides functionality expected by the user in any sort of execution circumstances and enables examination of other behaviors that may be associated with adaptability. The *Gather Information* choices include observing the environment with binoculars, viewing the fragmentary order (FRAGO) to reference the mission and tasks, viewing intelligence summaries (INTSUM) which will contain new information when appropriate to the vignette, and viewing the map to refresh one's memory of the terrain and laydown of forces. All four *Gather Information* icons are available on the interface during all phases of the simulation as it unfolds, as illustrated in Figure 4. Choosing to gather information pauses the simulation until the user chooses to resume. When the vignette resumes, the user enters the segment where he left off.

[1]The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.

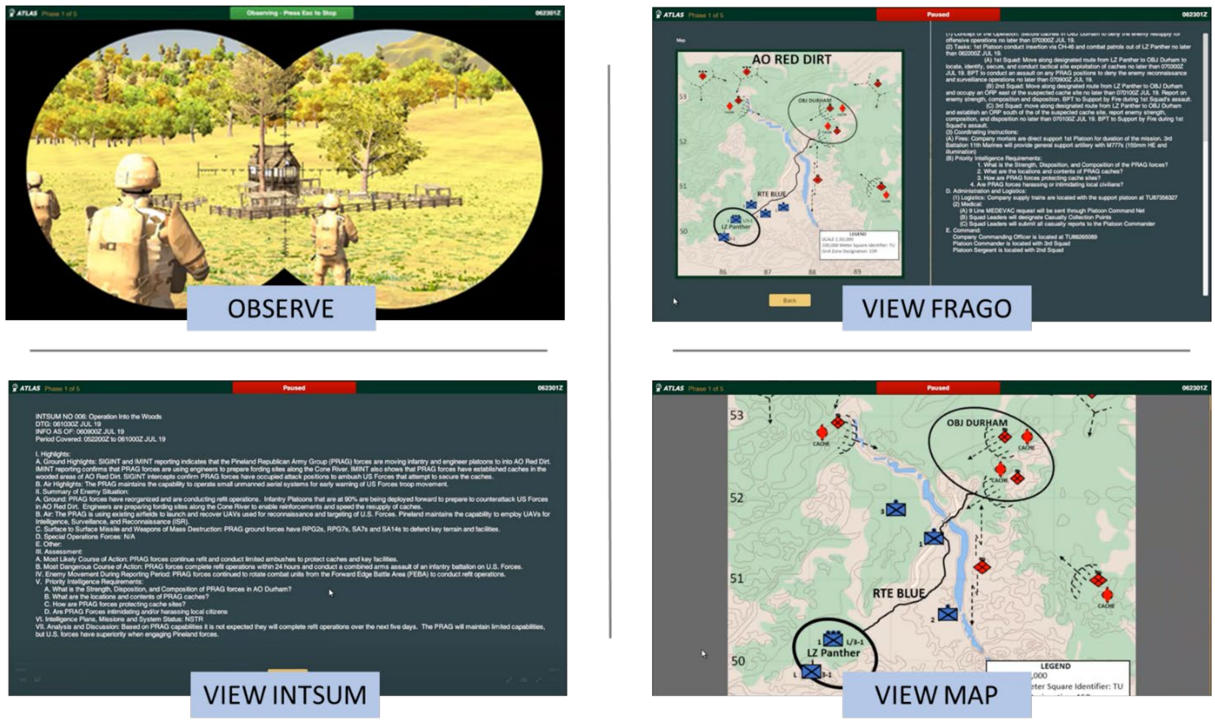

Figure 4. ATLAS *Gather Information* menu screenshots.

Adapt. The purpose of the *Adapt* response options is to change an aspect of mission execution as a result of a change to the situation. From a measurement perspective, the *Adapt* response options produce the time and quality measures of adaptability. Users can choose to adapt at any point during the simulation. The expectation is that users will adapt when they have recognized either the first, subtle, indicator of the trigger event, or the second, more noteworthy, level of the trigger event. Trigger events inject information that changes the situation in a way that impacts the success of the initial mission plan and therefore should elicit a change to the plan, or adaptation. The vignettes are designed such that eventually the decision to adapt will be forced upon the user by the third level of the trigger event, which forces action. Choosing to adapt stops the vignette segment. After the user has completed all desired adaptations, the simulation resumes by starting in the next vignette segment.

As described previously, user's adaptations to one vignette segment do not modify the next vignette segment. That is, if the user decides to alter the squad's formation or the positioning of the fire teams based on the events of Segment 1, those adjustments may or may not be present when Segment 2 begins. The rationale for this limit on users' interaction with the simulation is to present each user with the same information so that adaptive responses can be compared across test populations. Therefore, each new segment of the vignette is proceeded by a lead-in communication from the platoon commander with an order providing a justification for the array of friendly forces. In other cases, the next vignette segment is fast-forwarded in time sufficiently to justify the system ignoring the user's last set of adaptive decisions. This approach to measuring complex skills in a tactical context was validated in previous research, where a Decision Requirements Interview using paper and pencil scenarios was validated as a means of assessing squad leader decision skill (Ross et al., 2015).

As indicated earlier in Figure 3, the ATLAS system collects both adaptive response data and exploratory, behavioral pattern data. The logging system will record all interface interactions, 3D world events, and timings. The adaptive response data comes from the *Adapt* set of response options and is used to compute the five established measures. In addition to capturing those data, the system captures keystroke, button clicks, and time on menu pages. These data will allow us to examine and analyze behavioral patterns that were not hypothesized to be associated with adaptability in prior planning, but may be related to adaptive skill. For example, it may be that users who gather information with greater frequency demonstrate higher adaptability scores in other areas. Any behavioral patterns found to be meaningfully associated with adaptability will suggest additional, or replacement, measures of adaptability to

[1]The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.

instantiate in the final ATLAS system. These patterns would also be informative for creating assessment measures in other, more interactive training simulations.

Standard Item Structure

Each vignette phase contains an "item sequence" in which a challenge is presented to the user to elicit a response set that provides data used to calculate each of the five measures. An item sequence can be compared to a more traditional test item, such as a story problem. In our case, the story problem includes several questions following each story. The phase of the vignette is the story part of an ATLAS item sequence. For each vignette phase the user reacts to the trigger event by deciding when to adapt (the speed of reaction measure), assesses the new situation by deciding how priorities have changed due to the trigger event (the quality of situation assessment measure), and decides what adaptations to make (the quality of adaptive action measure). The clock is running as the user completes these responses, so duration measures are taken for length of time to re-prioritize concerns and length of time to decide how to adapt. Figure 5 provides a visual depiction of the item sequences present in every vignette phase, with each item sequence reading from left to right.

Users will receive a score for each stage of the cognitive process of adaptability, along with an overall score.

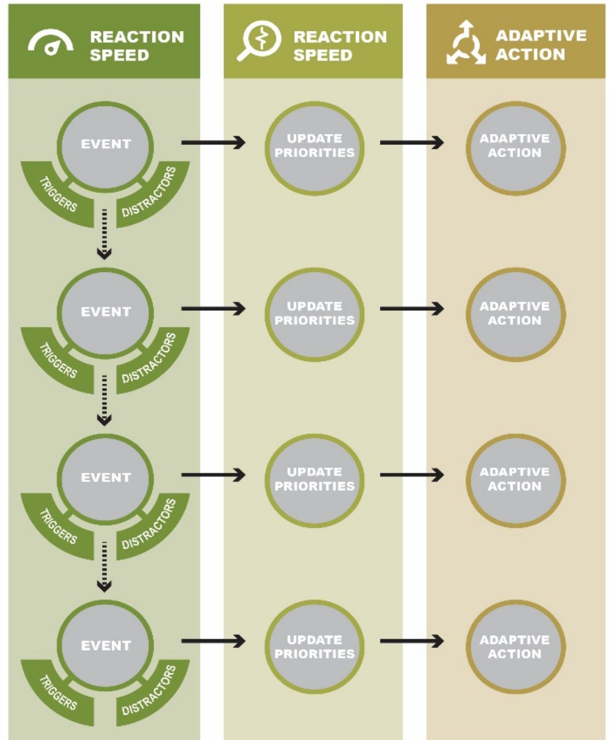

Figure 5. Item sequences require reaction to a trigger event, updating of priorities, and adaptive action.

USER TESTING

To date, the ATLAS prototype has been subjected to three rounds of user feedback sessions and testing. The primary goal of this pilot testing was to improve the immersivity and realism of the system, and to make the response interface more intuitive for Marines users.

The purpose of Pilot Test A was to collect feedback on early functionality and navigation flow of the skeletal framework of the ATLAS prototype. Pilot Test A used paper mockups of the initial ATLAS prototype to conduct a Wireframe Test in which feedback was gathered about page layout, interface elements, navigational systems, and functionality of the tool. Testing was conducted with four internal personnel and colleagues with appropriate military experience. Results indicated users interacted with the interface as anticipated and few recommended changes were identified.

Pilot Test B collected usability feedback to determine if the interface was intuitive for Marines, if the flow of the vignette made sense and kept the Marines engaged, and if the terminology used was appropriate. Participants in Pilot Test B were nine infantry Sergeants currently serving as instructors at The Basic School at Quantico, Virginia. All participants had a moderate level of squad leader proficiency (i.e., approximately 18 months served as a squad leader). During the two-hour testing session, Marines reviewed a tutorial presentation and then completed a think-aloud interview on the ATLAS demonstration and paper mockups of the full ATLAS prototype. Findings indicated that the interface and navigation allowed Marines to make decisions and take actions effectively and the flow of the vignette was intuitive. Specifically, participants felt that action options presented in the tool were a good representation of potential squad leader adaptations and that the tool was operationally relevant. Participants also felt that the tool was

[1]The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.

engaging and stated interest in utilizing the tool once it was fully developed. Participants provided minimal recommended changes to screen content and tactical aspects of the vignette.

The purpose of Pilot Test C was to collect performance data and user feedback required to prepare the prototype for its validation during field testing. Participants in Pilot Test C were nine "experts," ranked Sergeant to Mastery Gunnery Sergeant, (i.e., approximately 10 years of infantry squad leader and platoon sergeant experience to include time served as a mentor) and seven Marines, ranked Corporal to Gunnery Sergeant, who represented lower levels of infantry squad leader proficiency (i.e., less than 1 year of infantry squad leader experience). The two-hour test session incorporated a fully functional ATLAS vignette prototype with six item sequence, a demographic form, and a functionality and usability form. Findings suggested that the identified measures of squad leader adaptability were appropriate, and with minor adjustments to vignette structure and user interface, the ATLAS tool will more precisely capture data in a standardized fashion. Feedback indicated that use of the "on-rails" approach to achieve a balance between a naturalistic environment and standardized data capture was effective but required a more discernable presentation of vignette distractors and triggers to make the assessment item response points more well-defined.

CONCLUSIONS AND NEXT STEPS

The purpose of this effort is to design and develop a sustainable virtual simulation for assessing complex cognitive skills that is engaging and realistic while maintaining a standardized measurement structure with an automated scoring system. Efforts to date have resulted in the development of the system interface and user response functionality instantiating the automated assessment approach into a prototype vignette.

We identified a cost-effective virtual format for capturing standardized assessment of complex cognitive skills that removes the requirement for a human rater. The system interface design functionality is intuitive to users and meets expectations as to the options a typical Marine squad leader would have available during a mission. The visual presentation of the vignette and the user response interface engaged users and adequately simulated a naturalistic tactical environment. Maintaining an "on-rails" approach to the vignette design allowed for standardized data capture and did not distract from user engagement. Based on user feedback, minor adjustments will be made to the vignette format and user interface to improve the precision of the data capture.

The system prototype developed during this effort utilizes automated algorithms to calculate hypothesized measures of the cognitive process behind adaptability, which relieves the need for a human rater. However, the extent to which the human rater can be removed from the scoring process will be better understood following the next testing in which the validity of the hypothesized measurement algorithms will be examined with a diverse user group.

For basic sustainability of the tool, a repository of vignettes containing three matched vignette pairs has been identified. Each vignette pair differs in setting and mission, but all six vignettes are parallel in construction with respect to difficulty and length. These matched vignette pairs enhance sustainability of the tool and enable repeated assessment, where each test session assesses adaptability by presenting participants with one of the vignettes from each pair. Once the initial measurement approach has been validated, testing will occur to evaluate the equivalence of the vignette pairs.

In addition, to increase the sustainability of the tool further, we identified an approach to expanding the repository of scenarios by utilizing a Procedural Content Generator. The Procedural Content Generation function will automatically modify parameters of the developed vignettes including terrain features such as vegetation density and elevation variation, and environmental objects such as building, vehicles, and even civilians. Future phases of research will examine the validity of the measures when a rapid content generation function is applied.

The outcomes of this effort will contribute to the research community by providing a framework, method, and automated measures for assessing a complex cognitive skill in a virtual environment. This line of research will result in a validated assessment tool as well as a number of lessons learned for the research community. Questions will be answered with regard to whether or not the hypothesized cognitive framework holds up to scrutiny and therefore extends our understanding of the processes contributing to the construct of adaptability, the boundary conditions for rapid content generation using automated tools, and automated decision metrics that can be applied to other assessment tools and embedded as metrics in training tools.

[1]The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.

ACKNOWLEDGEMENTS

This work is funded by the Office of Naval Research, Human Performance, Training & Education Program. The authors would like to thank Dr. Peter Squire for his support and wise guidance; Colonel Greg Poland, Mr. Stu White, Dr. Alana Nicastro, Major PJ Tremblay, and the staff and Marines of Marine Corps Tactics and Operations Group for their significant insights and generous contribution of time and expertise; and The Basic School for their domain expertise and continued contribution to improving decision assessments. The views contained in this paper are those of the authors and do not reflect the perspectives of the Department of Defense, the military services, or any government research agency.

REFERENCES

Asymmetric Warfare Group (AWG) and Johns Hopkins University Applied Physics Laboratory (2016, October). Assessing Soldier Adaptability, Phase II: Metrics. Fort George Meade, MD: AWG.

Baard, S. K., Rench, T.A., & Kozlowski, S. W. J. (2014). Performance adaptation: A theoretical integration and review. *Journal of Management*, 40(1), pp. 48-99.

Boyd, J. R. (1987). *Organic design for command and control.* (Unpublished briefing). Retrieved from http://pogoarchives.org/m/dni/john_boyd_compendium/essence_of_winning_losing.pdf

Endsley, M. R. (2000). Direct measurement of situation awareness: Validity and use of SAGAT. In M. R. Endsley & D. J. Garland (Eds.), "Situation Awareness Analysis and Measurement" (pp. 147-173). Mahwah: Lawrence Erlbaum Associates.

Kozlowski, S. W. J., & Rench, T. (2009). Individual differences, adaptability, and adaptive performance: A conceptual analysis and research summary. Technical Report No. 08146, Battelle Scientific Services, Research Triangle Park, NC.

OSD P&R (2009) Adaptive FORCE Program manual, Developed by The Office of the Undersecretary of Defense (OSD) for Personnel and Readiness (P&R) for The Basic School, Quantico, VA, October.

Phillips, J. K., Ross, K.G., Rivera, I. D., & Knarr, K. A. (2013). Squad leader mastery: A model underlying cognitive readiness interventions. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference.* Available at http://www.iitsecdocs.com/search. Arlington VA: National Training Systems Association.

Ross., K. G. & Phillips, J. K. (2017). A cognitive framework and approach for adaptability assessment. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference.* Available at http://www.iitsecdocs.com/search. Arlington VA: National Training Systems Association.

Ross, K. G., Phillips, J. K., & Knarr, K. A. (2015). The small unit decision making assessment battery: Development and psychometric analysis. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference.* Available at http://www.iitsecdocs.com/search. Arlington VA: National Training Systems Association.

Ross, K. G., Phillips, J. K., & Rivera, I. D. (2013). Marine Corps Maneuver Squad Leader Mastery Model. Orlando, FL: Cognitive Performance Group.

[1]The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.