A Methodology for Simulation-based Job Performance Assessment

Sowmya Ramachandran, Jeremy Ludwig
Stottler Henke Associates, Inc.
San Mateo, CA
sowmya; ludwig @stottlerhenke.com

Eduardo Salas, Michael Rosen
University of Central Florida
Orlando, FL
esalas; mrosen@ist.ucf.edu

ABSTRACT

Job performance measurement is of critical importance to any organization's health. It is important not only to recognize and reward good performance, but also to groom future leaders. Developing effective assessment techniques that are valid, effective and fair is an ongoing challenge. Assessing factual knowledge using multiple-choice test batteries relatively inexpensive and tends to be commonly used. Hands-on assessment is the most effective in assessing task proficiency but is very resource intensive and expensive. Computer-based simulations provide an alternative where users can be assessed in the context of skill application under controlled conditions. However, simulations are expensive to produce and maintain. Validated guidelines and methodologies are needed to help organizations develop effective assessment simulations. In this paper we present a standard, prescriptive methodology for developing simulations for job performance assessment. We then describe a performance assessment simulation for Light-Wheeled Vehicle Maintenance constructed according to this methodology. This simulation includes automated assessment methods that borrow heavily from existing work in intelligent tutoring systems. Finally, we discuss future research directions based on the results of this initial methodology and assessment.

ABOUT THE AUTHORS

Dr. Sowmya Ramachandran is research scientist at Stottler Henke Associates, a small business dedicated to providing innovative Artificial Intelligence solutions to real-world problems. Dr. Ramachandran's interests focus on intelligent training and education technology including intelligent tutoring systems (ITSs) and intelligent synthetic agents for simulations. She has developed ITSs for a range of topics including reading comprehension, high-school Algebra, helicopter piloting, and healthcare domains. She has participated in workshops organized by the Learning Federation, a division of the Federation of American Scientists, to lay out a roadmap for critical future research and funding in the area of ITSs and virtual patient simulations. She has developed a general-purpose authoring framework for rapid development of ITSs, which is currently being used to develop an intelligent tutor training system Navy Tactical Action Officers. She has also developed tools and technologies for training emergency first responders.

Jeremy Ludwig is a Project Manager and Lead Software Engineer at Stottler Henke. He joined after completing his Master's Degree (2000) in Computer Science at the University of Pittsburgh with a concentration in Intelligent Systems. His research areas include intelligent training systems, behavior modeling, and machine learning.

Dr. Eduardo Salas is Trustee Chair and Professor of Psychology at the University of Central Florida. He also holds an appointment as Program Director for Human Systems Integration Research Department at the Institute for Simulation & Training. Previously, he was a senior research psychologist and Head of the Training Technology Development Branch of NAVAIR-Orlando for 15 years. During this period, Dr. Salas served as a principal investigator for numerous R&D programs focusing on teamwork, team training, advanced training technology, decision-making under stress, learning methodologies and performance assessment. His expertise includes helping organizations on how to foster teamwork, design and implement team training strategies, facilitate training effectiveness, manage decision making under stress, develop performance measurement tools, and design learning

environments. He is currently working on designing tools and techniques to minimize human errors in aviation, law enforcement and medical environments. He has consulted to a variety of manufacturing, pharmaceutical laboratories, industrial and governmental organizations. Dr. Salas is a Fellow of the American Psychological Association (SIOP and Division 21), the Human Factors and Ergonomics Society. He received his Ph.D. degree (1984) in industrial and organizational psychology from Old Dominion University.

Michael A. Rosen is a doctoral candidate in the Applied Experimental and Human Factors Psychology program at the University of Central Florida and has been a senior graduate research associate at the Institute for Simulation and Training since the fall of 2004 where he won the student researcher of the year in 2006. He is currently a MURI-SUMMIT graduate research fellow and focuses on developing theory, methods and tools for understanding and measuring cognitive and social processes in team problem solving. His research interests include individual and team decision-making and problem solving, human-computer interaction, performance measurement, and simulation-based training in high-stress high-stakes domains such as healthcare and the military. He has co-authored over twenty peer reviewed journal articles and book chapters related to these interests as well as numerous proceedings papers and presentations at national and international conferences.

A Methodology for Simulation-based Job Performance Assessment

Sowmya Ramachandran, Jeremy Ludwig
Stottler Henke Associates, Inc.
San Mateo, CA
sowmya; ludwig @stottlerhenke.com

Eduardo Salas
University of Central Florida
Orlando, FL
esalas@ist.ucf.edu

INTRODUCTION

The efficiency and effectiveness of an organization depends very crucially on its workforce. Job performance assessment carries high stakes for everyone involved. For employees, it determines their pay grades and promotions and thus plays a major role in their career advancement. For an organization, good performance assessment is crucial to its long-term health and sustainability. Given the stakes, fairness and validity of assessment are very important concerns.

There are several ways of assessing performance (Campbell et. al. 2004). A traditionally accepted approach is to use multiple-choice questions that have been carefully designed and validated. Situation Judgment Tests present cases or situations along with a set of possible actions. The examinee is expected to judge each of the choices and make an optimal choice. Such tests are used to assess judgment skills. Simulations of various types are also used for job performance assessment. Path simulations present limited interactivity where examinees are presented with simulation scenarios and several pre-defined paths to follow. The users' answers along the way determine the path they take. On the other hand, open simulations present users with a wider array of choices and their actions can change the state of the simulation. These types of open simulations offer more interactivity and power but are also more expensive to produce. Hands-on assessments observe examinees in a standardized operating environment as they perform tasks on real equipment. These assessments come the closest to testing on the job knowledge but are resource-intensive.

Each of the above approaches has its strengths. Multiple-choice questions are easier to develop and thus make it possible to cover a wide variety of skills relatively inexpensively. However, the problem of inert knowledge is well-known and well-documented (Schank 1995). Inert knowledge reflects the phenomena where people possess sufficient factual knowledge but lack the proficiency to apply this knowledge to solve real problems. For example, a light-wheeled vehicle mechanic may have knowledge of all parts of a HUMVEE and how they connect with each other, but may lack the practical skills for troubleshooting a defective vehicle efficiently. This is an example of inert knowledge.

Hands-on tests, on the other hand, are highly regarded within the Army for their validity. They do have several drawbacks. First, they require one-on-one time between the examinee and at least one assessor. Second, it is difficult to ensure fairness and objectivity in assessment in such settings. Often it is recommended to use two assessors to ensure objectivity but this leads to further increase in resource requirements.

Simulations provide many of the benefits of hands-on testing in that they assess skills in the context of a realistic work situation. Thus, they get around the problem of inert knowledge. Simulations typically include automated performance assessment. This overcomes the issues to uniformity and objectivity and eliminates the need for one-on-one time with an assessor. However, simulations are much more expensive to develop than multiple-choice batteries. Furthermore, ensuring validity is a challenge. There are no guidelines for developing them. An assessment simulation must measure relevant skills and must be valid. Care must be taken to ensure that the simulations measure job skills and not the ability to use computers or the ability to game the system.

In this paper we present a standard, prescriptive methodology for developing simulations for job performance assessment. We then describe a performance assessment simulation for Light-Wheeled Vehicle Maintenance constructed according to this methodology.

SIMULATION-BASED ASSESSMENT DEVELOPMENT METHODOLOGY

The methodology development process was driven by 1) a review of current literature on the design of simulation scenarios and measurement tools as well as the development of selection systems and test items, and 2) practical experience implementing the methodology in developing the prototype simulation.

This section details the eight-step methodology we developed while creating an initial assessment scenario.

Step 1. Define clearly what needs to be measured

Any effective measurement system begins with a clear definition of what is to be measured. In this case of using performance in simulations as an indicator of performance, the ultimate goal is to obtain a measure of proficiency in the knowledge, skills, aptitudes, and other characteristics (KSAOs) underlying effective performance within a domain. The tasks to be performed in this step include:

- Perform a document review of pre-existing job-analysis, training materials, technical manuals, and standard operating procedures. These documents are often readily available in the military where the competencies for jobs have been clearly articulated.
- Conduct structured interviews with SMEs.
- Compile a list of competencies and associated performance contexts.

Step 2. Develop a sampling strategy

To ensure that the entire domain (or critical aspects of the domain) are represented in the test—the simulation scenarios and events—a strategy for developing scenarios, events and critical responses must be developed that meets two high level goals. First, each scenario including the events and targeted responses must be clearly linked to the targeted competencies. This ensures that aspects of performance not related to the domain competencies do not become a part of performance measurement and subsequently the selection decision. This reduces the level of construct contamination in the measure. Second, systematically linking scenario development to the targeted competencies affords the ability to track what competencies have and have not been sampled by the simulation scenario. This ensures the opportunity to sample the entire domain and to avoid under-representing (or under-specifying) the targeted competencies in the performance that the simulation captures.

When test length is an important concern, sampling the competencies that are most discriminative is a logical strategy. Additionally, methods of sampling strategies for competencies can focus on time, criticality, and level (Sackett & Laczo, 2003). That is, competencies can be chosen based on the relative amount of time individuals spend on the job using the specific competency, the degree to which the competency distinguishes between successful or unsuccessful staff, or the degree of the competency needed to perform successfully on the job.

An idealized approach would involve the following steps if SME ratings of frequency, criticality, difficulty, and level of activity and knowledge focused competencies are not already available. First, the results of step one of this process would be used to develop a survey. This would be distributed to SMEs for a given Military Occupational Specialty (MOS) and contain items for each identified activity-focused competency, performance context (i.e., more specific instance of an activity competency), and knowledge-based competencies. SMEs would provide ratings of difficulty, criticality, frequency and level. This data based could then be used to sample a range of competencies for the construction of an individual scenario as well as for constructing sets of multiple scenarios to be used as alternative forms (i.e., these SME ratings can be used as initial validity evidence that two sets of scenarios sample equivalent competencies).

Steps 3. Generate scenarios with embedded events and measurement tools

The process of developing simulation scenarios is central to using simulations for selection purposes. Cognitive and behavioral task analysis techniques (e.g., critical decision method, hierarchical task analysis) can be leveraged to sample the range of tasks required and situations encountered for a specific job. The Critical Decision Method and other event-based knowledge elicitation techniques can be used to generate critical events and targeted responses that can be linked to the competencies of the domain. For procedural skills, the fundamental outlines of simulation scenarios can often be generated from existing technical and training references.

Once an outline of the simulation has been created, the general process involves progressively contextualizing the abstract competencies, using SME guidance to focus on key competencies, using supporting documentation to generate the overall structure of a scenario, and using SME interviews to provide details about each component of the procedural task. The end goal of this process is to create a simulation scenario and populate it with 'items' (i.e., the scenario events) to which the user is expected to respond. Scenario events should be realistic, aim at the appropriate level of difficulty, provide multiple opportunities to display targeted competencies, and sequential dependencies

should be avoided in the measurement associated with events (Fowlkes & Burke, 2005).

Step 4. Decide on an appropriate scaling technique and encode in a measurement tool

The nature of responses to simulation events is critical in determining the correct scaling technique. For this reason, the scenarios need to be created before determining how to assess the scenario responses.

There are multiple ways to capture performance in simulations. Event-based measurement can result in dichotomous scoring (e.g., did the individual exhibit the targeted behavior?) or through other types of ratings (e.g., Likert type scaling in Behavioral Observation and Behaviorally Anchored Rating scales). Deciding on the best scaling technique involves considering the characteristics of the performance being measured as well as the goals of the measurement system (in this case, selection).

In terms of scaling methods for performance, common metrics include either 1) latency from the time some information is provided to the performance of an expected action, 2) a dichotomous scoring of whether an action was or was not taken, or 3) a count of 'missteps' before performing the targeted response. All three of these are possible for most items and are likely useful in any procedural skill task. It is likely that the dichotomous scoring is the most straightforward and easy to interpret in most cases; however, the number of missteps and latency measures are likely more diagnostic between different skill levels. Dichotomous scoring is likely to give the simplest measure of basic competence while the other approaches are more likely to distinguish between competence levels at finer levels of detail.

Step 5. Have scenarios reviewed by subject matter experts (SMEs)

Just as it is recommended for SMEs to review test items during development of traditional selection tools, SMEs can provide valuable insight into how representative the scenarios and measurement tools are of actual performance. This relatively simple step ensures the 'face validity' of the scenarios, a facet that can greatly affect how an individual perceives and performs within the simulation. It also serves as a check on the appropriateness of the sampling strategy developed and implemented.

Step 6. Administer the simulation and measurement tools to a developmental sample

The simulation should be run with a sample from the intended population of use for validation purposes. Additionally, measurement of this sample's subsequent performance on the job should be collected. This data will allow for validation and optimization of the simulation test.

Step 7. Evaluate the scenarios and measurement tools

Using the data from the developmental sample, the characteristics of the simulations scenarios and measurement tools can be evaluated. Specifically, the item response characteristics for each scenario event can be determined. This will enable the process of choosing and refining the simulation test to meet the specific requirements of the selection task.

The primary means by which this is accomplished is through correlating simulation scores with other measures of competency. However, additional work is required to establish the validity of multiple sets of scenarios as equivalent tests of competency. This problem is equivalent to developing parallel forms of tests in traditional test or selection tool development. There are several options available to establish the validity of using parallel or alternate forms of tests (in this case, different sets of simulation scenarios). The first strategy (which is likely the strongest) involves administering both sets of scenarios to the same group of individuals. Ideally this group of individuals would represent a continuum of competency (e.g., people from different skill levels, different levels of tenure, a wide distribution of on-the-job performance scores, etc) so that there is variation in the scenario scores between participants. The degree to which the individual's scores on the different scenarios are correlated is evidence of the validity of using the scenarios as equivalent tests. Second, scenario scores from each set can be correlated with other measures of competency (e.g., knowledge tests, situational judgment tests, supervisor ratings, groups of expert and novice test takers, etc.). This can be done in conjunction with the first strategy or in a between subjects fashion with each set of scenarios being administered to separate groups. The degree to which the two sets of scenarios show similar patterns of relationships with these other indicators of competency can be taken as evidence of the validity of using the two sets of scenarios as parallel test forms. Third, the scenarios can be reviewed in terms of the degree to which they reflect or sample the same competencies. This review would involve subject

matter expert ratings of the criticality, frequency, and difficulty of the activity competencies, domains of knowledge, and contexts of performance reflected in each set of scenarios. The degree to which these ratings match is evidence of the validity of using the two sets as parallel forms. All of these strategies can be employed to build the strongest case possible for using different sets of scenarios as equivalent.

Step 8. Optimize the selection test

The simulation-based test can be optimized using information from the evaluation of the data gained from the developmental sample. This information can be used to maximize the predictive power of the test (e.g., increase reliability of measurement at the chosen criterion cutoff; increase diagnosticity over ranges of proficiency as needed). As in traditional scale development, test length and predictive power of the test are often at odds with the practical considerations demanding the shortest tests possible. This is the case with simulations as well; using item response theory and psychometric principles of test design, the shortest tests (simulations) can be designed with the highest level of prediction and therefore the most utility in selection.

METHODOLOGY IN PRACTICE

In this section we describe an example assessment scenario created using the methodology described above. The 63B mechanic MOS was selected as the target for developing a prototype assessment scenario.

Step 1. Define clearly what needs to be measured

Step 1 began with a review of the available documentation on the 63B MOS. This included prior and available job analyses, technical manuals, standard operating procedures (SOPs), and training materials. The core competencies for the 63B MOS were adopted from existing Army documentation:

- Preventive Maintenance Checks and Services (PMCS)
- Perform scheduled maintenance tasks to keep vehicles operational
- Troubleshoot Vehicle and Equipment Problems
- Inspect and test equipment and determine the causes of malfunctions
- Repair Vehicles and Equipment
- Remove and replace components and to complete all necessary repairs, adjustments, and checks to make vehicles and other equipment operational
- Use Technical References
- Use resources and references in performing maintenance procedures
- Safety Procedures
- Follow safety procedures
- Be alert to possible dangerous or hazardous situations and take steps to protect self, other Soldiers, and equipment

The competencies listed above are activity focused (i.e., descriptions of tasks performed on the jobs) and not person focused (i.e., descriptions of the KSAOs required for performing the task). This is beneficial for developing simulations as the scenarios must provide opportunities to perform these activities. In addition to these activity focused competencies, several lists of tasks and a 'competency-based blueprint' were available (Moriarty & Knapp, 2007). This competency-based blueprint consisted of hierarchically organized knowledge categories (e.g., engines, electrical systems) involved in successful performance for 63B mechanics. When combined with the activity focused competencies, this provided a type of two-level competency framework. That is, to create the entire competency space for the 63B, it is necessary to cross the activity focused competencies listed above with the blueprint categories of knowledge (Moriarty & Knapp, p. 21). For example, troubleshooting (and activity focused competency) can be done within engines (and subsequently within gasoline and diesel fuel systems) and electrical systems (and subsequently within charging systems or task relating to basic principles of electricity).

Specifications of competencies for an MOS detail the 'what' but not the 'how' of performance. Since performance in simulation scenarios is dynamic and the specification of competencies is necessarily abstract there is a need for an intermediate step between competency and dynamic performance to help guide later steps in the process. This is analogous to defining specific learning objectives in the context of simulation-based training (SBT; Fowlke, Dywer, Oser, & Salas, 1998). Essentially, this involves generating 'performance contexts' associated with competencies—more specific and detailed descriptions of performance than those provided by the abstract activity focused competencies. For example, the 63B competency of 'Inspect and test equipment and determine the causes of malfunctions' was identified as crucial by SMEs; however, in order to generate scenarios that tap this competency, it was necessary to understand the contexts of performance where these competencies would be displayed. Based on SME interviews, electrical and hydraulic systems were identified as

general areas where diagnostic skills were most vital. Further interviews provided more information about specific cases where diagnostic skills could be evaluated. Additionally, the 63B competency of 'use resources and references in performing maintenance procedures' was further contextualized with SME input to the interpretation and use of schematics. This was identified by SMEs as a means for distinguishing between skill level 1 and 2 mechanics.

Step 2. Develop a sampling strategy

Step 2 was limited in this case since only a prototype system was being developed. The prototype is designed to assess the following skills:

- KSA 1: Troubleshoot vehicle and equipment problems
    - Inspect and test equipment and determine the causes of malfunctions
- KSA 2: Use technical references
    - Use resources and references in performing maintenance procedures

Steps 3-4. Generate scenarios with embedded events Decide on an appropriate scaling techniques and measurement tools

Steps 3 and 4 were completed for these two skills in creating the prototype system. The resulting storyboard contains simulation events, targeted responses and measurement approaches. Because the 63B mechanics tasks are highly proceduralized, a minimal amount of further cognitive or behavioral task analysis was required. In the case of the 63B MOS, scenario events were defined primarily in terms of information provided to the mechanic from the vehicle or through the various tools available to the mechanic. Events were defined in terms of the action the mechanic should take given the provision of this information (i.e., the event). These were modeled on the troubleshooting guides obtained from training materials and the SME review.

Table 1 Initial portion of prototype scenario.

	Event	Targeted response	Additional information	Possible metrics
NA	Mechanic is provided with 5988-E form detailing problems and history of vehicle (*example provided in separate document*). (KSA2)	Mechanic selects appropriate technical reference (electrical system for HMMWV M998) from the sources available.	-There are different types of HMMWV' s. The procedures outlined in this scenario are for the M998 (the basic model). -The major distracting information in this step involves 1) sections of the manuals for other types of HMMWV or other trucks (e.g, if the mechanic selects information on M1044A1, they have not been able to extract the appropriate information from the 5988-E form), and 2) sections of the appropriate manual (i.e., the M998) that	-time from presentation of the 5988-E to accessing the correct troubleshooting procedure -dichotomously scored (mechanic did or did not access correct reference) -number of incorrect references

			do not match the specific problem (i.e., not the appropriate troubleshooting procedures—e.g., in this case something other than the electrical system).	accessed
Step 1	Mechanic accesses correct troubleshooting procedure within the technical manual (*possibility: if after specified amount of time, mechanic does not locate this procedure, he/she is cued to do so*). (KSA2)	Mechanic tests the specific gravity of the electrolyte in the battery using a battery tester. -In the field, testing the battery involves taking a drop of the battery fluid and placing it in a battery testing device (a hydrometer); the mechanic then holds the device up to the light and looks through an eyepiece; the reading shows up as a horizontal line on a scale.	-needs hydrometer (battery tester) or equivalent information (information can be provided with a simple line and scale). -there are two batteries in the M998, each with six separate cells that must be tested individually; if any one cell' s specific gravity is below 1.250, the entire battery must be replaced. There is a figure of the battery in Vol 2 of the TM, section 4.79. The TM cites a different TM (9-6140-200-14) which we do not have for more details on this process.	-time from accessing correct troubleshooting procedure to testing specific gravity of battery -dichotomously scored (mechanic did or did not test all battery cells)
Step 1-2	Mechanic is provided with informatio	Mechanic removes companion seat/battery box Removes and	-the battery compartment is located under the passenger seat; there is a figure and outline of process to remove the companion seat/battery	-time from receiving information that batteries' specific gravity is

	n that the specific gravity of the battery is above 1.250 (KSA1)	cleans all battery cable connections	box in Vol 3 of TM; section 10-35. -SME' s said that battery cables would be visually inspected; if they looked corroded they would be removed and cleaned.	ok to removing battery box and cables. -dichotomous scoring (yes or no, did mechanic remove battery box and cables)
Step 2	Multimeter interface is presented to mechanic with improper settings (KSA1)	Mechanic sets multimeter correctly (to ohms) Tests for continuity across the shunt in the battery compartment; to test for continuity, the mechanic must locate the shunt using the schematic, and place a probe on the connection entering and leaving the shunt—see figure 2.	-the correct setting for the mutli-meter is ohms for this part of the task; the multimeter (AN/PSM-45) interface is detailed in TM-6625-3052-14 -major ' bad moves' for responses to this event include 1) setting the multimeter incorrectly, and 2) placing the multimeter probes incorrect positions (that is, the mechanic has to be able to read the schematic correct – figure 2 in troubleshooting procedures—in order to find and test the shunt).	-latency: time from accessing multimeter to 1) adjusting setting, and 2) checking for continuity. -dichotomous scoring: did the mechanic check the appropriate connection

Step 5. Have scenarios reviewed by subject matter experts (SMEs)

The initial scenario framework and evaluation tools were reviewed by mechanics during a Ft. Jackson site visit. SMEs were walked through the scenario on paper and asked to comment on each step as well as the scenario in general. Feedback from these SME interview/focus groups was used to add more contextual detail to the scenario and validate the accuracy and difficulty level of the scenario framework. Due to the highly procedural nature of the task, there were minimal modifications to the basic scenario.

Step 6-8

These steps were not carried out during the development of the prototype performance assessment scenario.

PROTOTYPE SIMULATION

We implemented the portion of the scenario described in Table 1 using the SimVentive tool for simulation

construction (Ludwig, Houlette, & Fu 2008). The simulation represents the scene as a series of html-based text and dynamic image maps (Figure 1). The user can view the vehicle from different angles, where image hotspots let the user perform various actions on the vehicle. For example, the user can click on a hotspot on the passenger-side door of the vehicle to get to its interior (as shown in the figure). The user can also refer to manuals and forms on the right-hand side panel by clicking on the hyperlinks. The simulator monitors the user's references to the technical manuals and forms as a part of its assessment. There are also tools that the user can select for various actions, where the simulation can assess the right tool usages and settings. For example, the user cannot use a multimeter to measure the specific gravity of battery cells. The simulation would mark this as an incorrect action. In addition, the tools must have the appropriate configuration for an action. For instance, the multimeter must be set to measure Ohms before the user can check across the shunt for continuity.

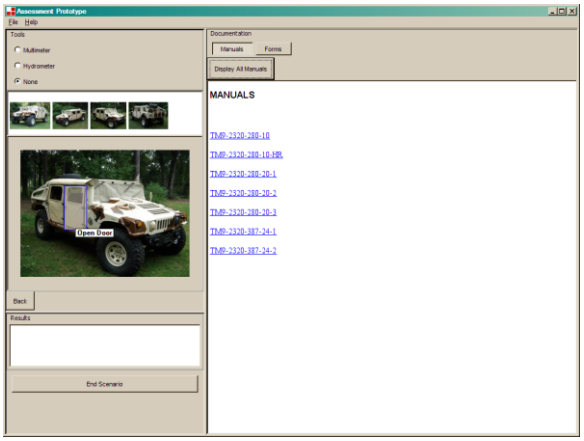

Figure 1: Prototype assessment simulation.

Figure 2 shows some additional aspects of the prototype simulation. First, the right-hand panel displays a reference manual showing a troubleshooting guideline. It also shows the simulations response to user actions on the lower left-hand side.

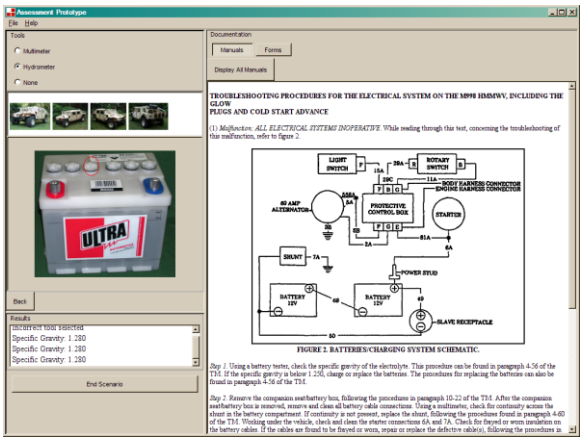

Figure 2: Prototype simulation interactions.

The simulation assesses performance based on a solution template approach designed for intelligent tutoring systems (Ong & Noneman, 2000). Figure 3 shows the template for the portion of the prototype scenario that was implemented. This template (shown as a tree in the authoring tool) specifies the procedure that the user must follow in this scenario. The bottom-level nodes in the tree are the direct actions that must be performed in the scenario. The interior nodes are task groups. The groups labeled with shaded boxes indicate that the actions in the group can be performed in any order (flexible ordering). The groups labeled with clear boxes indicate that the actions must be performed in order. Additional information about each action is specified in the right hand panel including an optional association with a KSA (labeled "principle"). The "Reason" field allows the author to specify an explanation to be shown during an optional debrief.

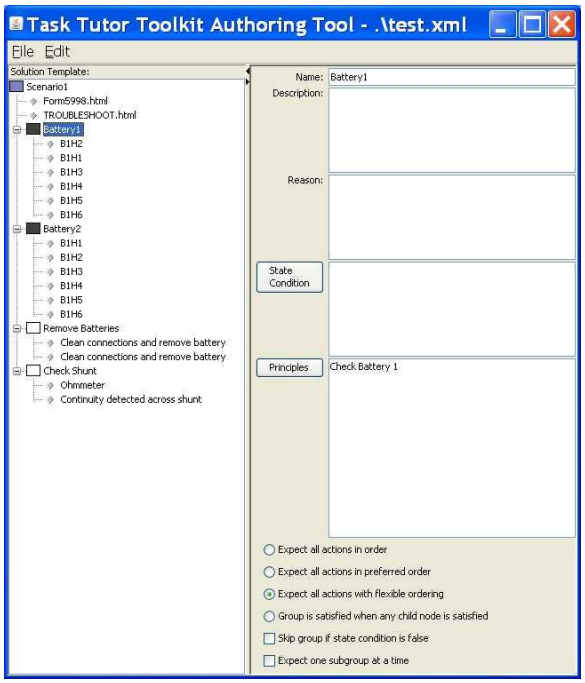

Figure 3: Solution template for the prototype scenario.

The simulator compares the user's actions with this template to assess his performance. An example assessment produced in the prototype scenario is shown in Figure 4. The overall score is arrived at by examining the appropriate actions completed in the preferred order (*Correct*), the appropriate actions completed out of the preferred order (*OK)*, and any actions that were not included in the solution template (*Unexpected)*.

Once the scenario was defined, it took about 40 man hours to develop the prototype implementation (which covers ½ of the scenario). Realistically, we expect that developing a completed assessment simulation end-to-end will take about four man-weeks. When amortized over the number of times it will be used, the cost for developing a simulation scenario is very small when compared to the cost of conducting hands-on job assessments with human facilitators and role players.

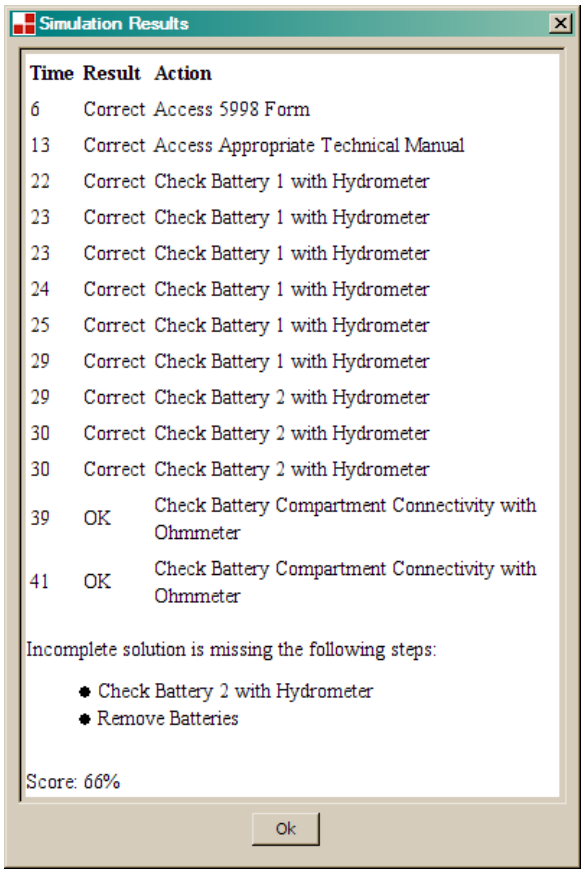

Figure 4: Example scenario assessment

CONCLUSION

The research described in this paper captures our initial efforts at creating a methodology for developing simulation-based assessments and building a set of simulation construction and assessment tools to support this methodology. Our initial feasibility study and prototype development has demonstrated that the theoretical framework for simulation development methodology can be implemented realistically and cost effectively in the real-world. Our future work in this area focuses on two main objectives.

The first objective is to develop a process that can be reproduced consistently to yield valid tests that will reliably and accurately measure skill levels. The methodology should provide enough guidance to enable Army personnel to develop such simulations with limited outside support. While the methodology presented in this paper is a step in the right direction, there is still a significant amount of work to do in this area. We plan to validate the methodology by using it to develop two assessment simulations in two distinct domains. This will demonstrate that the methodology is

practical, provide data on the effort involved in implementing the steps, and help refine it.

The second objective is to develop tools that will enable rapid development of assessment simulations. Cost is an important criterion determining the success of this line of research. Simulations are significantly more complex than current multiple-choice based assessments. In order to be competitive with them, simulation-based assessments should not only be demonstrably more effective, but also be cost-efficient. We plan to extend an existing simulation authoring tool to achieve this objective, focusing on simplifying the types of tasks commonly used in creating job performance simulations. Additionally, the extended authoring tool will also contain support for easily defining the performance assessment component of the simulation. The goal is to create end products that the Army can use with its own resources.

ACKNOWLEDGEMENTS

This work was funded under a contract from US Army Research Institute, contract number W91WAW-08-P-0070.

REFERENCES

Campbell, R.C., Keenan, P.A., Moriarty, K.O., Knapp, D.J., & Heffner, T.S. (2004). The Army PerformM21 Demonstration Competency Assessment Program Development Report (Technical Report 1152). Arlington, VA: U.S. Army Research Institute for Behavioral and Social Sciences.

Fowlkes, J. E., & Burke, C.S. (2005). Event-based approach to training (EBAT). In N. Stanton, Hedge, A., Brookhuis, K., Salas, E., & Hendrick, H. (Ed.), *Handbook of human factors and ergonomics methods* (pp. 47-41 - 47-45). Boca Raton, FL: CRC Press.

Ludwig, J., Houlette, R., & Fu, D.,(2008). Rapid simulation construction. Paper presented at the *2008 IEEE Aerospace Conference*, Big Sky, MT.

Moriarty, K. O., & Knapp, D. J. (2007). *Army enlisted personnel competency assessment program: Phase III pilot tests* (No. 1198). Alexandria, VA: United States Army Research Institute for the Behavioral and Social Sciences.

Ong, J., S. Noneman (2000). Intelligent tutoring systems for procedural task training of remote payload operations at NASA. Proceedings of the *Industry/Interservice, Training, Simulation & Education Conference (I/ITSEC 2000).*

Sackett, P. R., & Laczo, R. M. (2003). Job and work analysis. In W. C. Borman, D. R. Ilgen & R. J. Klimoski (Eds.), *Handbook of Psychology: Industrial and Organizational Psychology* (Vol. 12, pp. 21-37). Hoboken, NJ: John Wiely & Sons, Inc.

Schank, R. (1995), *What We Learn When We Learn by Doing*, Technical Report no. 60, Institute of Learning Sciences, Illinois.