Development of a Performance Assessment System for Combat Medic/Lifesaver Training

Roger M. Hamilton
National Research Council Research Associateship Program
Orlando, Florida
roger.m.hamilton@us.army.mil

Jessie Y.C. Chen
Army Research Laboratory
Orlando, Florida
jessie.chen@us.army.mil

Jack E. Norfleet
Army Research, Development and Engineering Command
Orlando, Florida
jack.norfleet@us.army.mil

John J. Anton
IVIR, Inc
University Park, Florida
janton@ivirinc.com

ABSTRACT

This paper documents the design, construction, verification, and validation of an automated and semi-automated hands-on summative performance assessment system intended to evaluate battlefield first responder skills (e.g., applying a tourniquet). The goal of this effort is to provide valid and reliable Army-standardized and unit-flexible Combat Lifesaver and Medic skills assessments designed to maximize automation while simultaneously minimizing administration demands and costs. The assessment system will be administered at U.S. Army Medical Stimulation Training Centers (MSTCs). MSTCs are standardized medical training platforms which enable classroom and hands-on simulation-based training of medical first responder skills. MSTC technologies provide the ability to control stimuli impacting the trainee as well as capture trainee performance data. The performance assessment system under development is part of the Medical Training Evaluation and Review System (MeTER) MSTC subsystem and will use testing rooms, programmable patient simulators and scenarios, and evaluator data collection systems (e.g., on desktop or tablet PCs) to provide standardized scenario-based (semi-)automated performance assessments. XML-based Simulations for Integrated Learning Environments (SIMILE^{TM}) software will integrate performance data automatically collected from simulation technologies, evaluator task performance scoring, standard and flexible scoring rubrics, and reporting requirements. Our test construction approach includes review of training documents, observation of current methods for combat medic hands-on skills assessments, and obtaining subject matter expert (SME) input to determine the assessment content and dimensions; scoring rubric development, and; verification and validation of both test and scoring rubric through piloting and SME review. Benefits of summative standardized assessments include Army-wide competence measurement, inter-unit comparisons, inter-Soldier comparisons, and providing feedback for training processes. Beyond MSTCs and medical tasks, however, the lessons learned here may help strengthen any performance assessment effort. Most importantly, MSTC-based assessments of first responder skills will inform training and operational decisions that could minimize preventable deaths on the battlefield.

ABOUT THE AUTHORS

Roger M. Hamilton is a Research Associate with the National Research Council, National Academy of Sciences. He has degrees from the United States Military Academy, the University of Southern California, and the University of Central Florida. He is currently working with the U.S. Army Research Laboratory at the Simulation and Training Technology Center in Orlando, Florida. His research interest lies in optimal learning environment generation.

Jessie Y. C. Chen is a research psychologist with U. S. Army Research Laboratory - Human Research & Engineering Directorate (field element in Orlando, Florida). Prior to joining ARL, she was a post-doctoral fellow with U.S. Army Research Institute for Behavioral and Social Sciences - Simulator Systems Research Unit. Her research interests include human-system integration, human-robot-interaction, vehicle crewstation design, and

medical simulation. She received her Ph.D. in applied experimental and human factors psychology from University of Central Florida in 2000.

Jack Norfleet is a Chief Engineer in the Soldier Simulation Environments division at the Army's RDECOM-STTC. Currently, he is in responsible for the medical simulation research efforts at the STTC. Mr. Norfleet has 24 years of experience in developing military simulations for training medical skills, and force on force skills. He has also worked on the development of various range instrumentation systems. Mr. Norfleet has a BSEE from UCF and an MBA from Webster University. He has also trained as an EMT. He is currently enrolled in the Modeling and Simulation PhD program at UCF.

John J. Anton is the Principal Investigator for Information Visualization and Innovative Research Inc. in Sarasota, Florida. He is responsible for conducting independent test and evaluation for medical simulation, learning management design and assessment for medical education, and productization services for medical education simulators and devices for both commercial and government customers. His current research areas include performance modeling, partial task trainer design, independent test and evaluation of live, virtual, and constructive simulations and models, live tissue replacement for medical education, modeling and simulation applied to autism, and medical education learning management.

Development of a Performance Assessment System for Combat Medic/Lifesaver Training

Roger M. Hamilton
National Research Council Research Associateship Program
Orlando, Florida
roger.m.hamilton@us.army.mil

Jessie Y.C. Chen
Army Research Laboratory
Orlando, Florida
jessie.chen@us.army.mil

Jack Norfleet
Army Research, Development and Engineering Command
Orlando, Florida
jack.norfleet@us.army.mil

John J. Anton
IVIR, Inc
University Park, Florida
janton@ivirinc.com

INTRODUCTION

The U.S. Army has strong human, operational, and financial interests in minimizing the effects of battlefield trauma. The skill level of battlefield first responders to this trauma, however, cannot be systematically determined because no standardized skill measurement program exists. This effort was a research and development program to provide prototype standardized Combat Medic and the Combat Lifesaver (CLS) hands-on performance assessments at Medical Simulation Training Centers (MSTCs) Army-wide. Because standardized assessments for these Soldiers are minimal, the "necessary guarantee to the People and our government that Soldiers, leaders and units can perform as designed and expected" (Asymmetric Warfare Group, 2009, p.7) may not exist, and the many other benefits of standardized assessments are not realized. The non-standardized and informal assessments currently in place can lead to the following problems: (1) the content, validity, reliability, and usefulness of the current informal assessments are unknown, so their true value to decision makers is unknown. For example, some important tasks are not assessed and tasks are assessed differently; (2) valid feedback for optimizing training processes is not possible, and; (3) potential cost and administrative savings are not realized.

MSTCs are standardized medical training platforms currently being fielded that enable classroom and hands-on simulation-based training of medical first responder skills. MSTC technologies provide the ability to control stimuli impacting the trainee as well as capture trainee performance data. The performance assessment system under development is part of the Medical Training Evaluation and Review System (MeTER) MSTC subsystem and will use testing rooms, programmable patient simulators and scenarios, and evaluator data collection systems (e.g., on desktop or tablet PCs) to provide standardized scenario-based (semi-)automated performance assessments.

The standardized advanced technologies found in the MSTCs can measure the situationally-identical competence of battlefield first responders. This potential can be realized through the systematic design, construction, verification, and validation of formal standardized assessments while also minimizing their cost, administrative demands and maintenance demands. These assessments can provide valuable feedback which informs schoolhouse, unit, and individual training and learning decisions that result in a more capable fighting force.

DESIGN

The assessment designs were to meet the following objectives: (1) construct to be assessed: current level of individual demonstrated expertise compared to standard of CLS and Medic curricula in minimizing the impact of battlefield trauma; (2) include a pre-test and a post-test; (3) be standardized *and* flexible; (4) be compatible with current MSTC technology; (5) maximize current automation and moves toward full automation; (6) support new technology impact measurement; (7) have a ratio scaled zero-to-infinity scoring system, and; (8) align with a cognitive assessment of the same material being independently developed.

Design Goals

Every performance assessment consists of: (1) an administration process; (2) the performance tasks to be assessed through the approved solution steps, and; (3) a

scoring rubric. A performance assessment should also have the following design goals for constructing those components: (1) maximize validity; (2) maximize reliability; (3) maximize training utility; (4) minimize administration demands, and; (5) minimize maintenance demands.

Maximize Validity

Validation is "the fundamental requirement in assessment development. Its importance cannot be overemphasized." (Bewley, Chung, Delacruz, & Baker, 2009, p. 302) Validity represents an overall judgment of the degree to which a construct has been translated into an operationalization (Trochim, 2009). In this case, how well do the assessments truly reflect the demonstrated necessary expertise of first responders compared to some standard?

The general goal is to build a body of evidence supporting a claim of validity. A strong claim of validity must address construct validity, content validity, criterion-related validity (e.g., predictive validity, concurrent validity, convergent validity, discriminant validity), and face validity. For example, for convergent validity, does the medic performance assessment score have high correlation with other indicia such as the cognitive assessment score, the Noncommissioned Officer Efficiency Review, the National Registry for Emergency Medical Technicians - Basic test score, Expert Field Medical Badge award, and Combat Medical Badge award? For face validity, is the assessment viewed by Soldiers as relevant, fair, and justifiably consequential? Being intensely systematic about validity would give decision makers and trainees great confidence that what they think is being assessed is real.

Maximize Reliability

Reliability is "the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials" (Carmines & Zeller, 1979, p.11). The intent of reliability is to "achieve sufficient score consistency or stability to warrant the use of scores in decision making" (Messick, 2000, p.3). Reliability is a prerequisite of validity (Freeman, Stacey, & Olivares, 2009). Reliability problems stem from sources of measurement error such as the lack of agreement among tasks and raters. The goal is to formalize and scientifically strengthen the assessment process by minimizing sources of assessment variability.

In a normal progression, MSTC sites were put on the ground rapidly to answer wartime needs and now need to be institutionalized. Identifying and mitigating current informal assessment variability is a logical next step in preparing MSTCs for formal assessments. Although some MSTCs have attempted to standardize locally, analysis of existing Army-wide documentation (Milham, Carroll, Stanney, & Becker, 2009) and observation of several MSTCs revealed several assessment variability sources as shown in Table 1.

Table 1. Sources of Variability in Current Assessment Processes

Variability Source	Current System
Administration process	
Administrative instructions (e.g., timing, content and tone)	Minimal standard
Equipment available to trainee	Minimal standard
Medical scenario setup	Minimal standard
Environmental stressors (e.g., simulated combat stimuli)	Minimal standard
Number of simulated casualties	Minimal standard
Number, location and severity of wounds	Not addressed
Reaction of simulated casualties	Minimal standard
Performance tasks	
Task inclusion	Minimal standard
Step clarity	Not addressed
Scoring rubric	
Evaluator step scoring	Minimal standard
Weighting of the tasks	Not addressed
Weighting of the steps	Minimal standard
Weighting of the repetition	Not addressed

Such variability must be mitigated in order to make a strong reliability claim. For example, in order to standardize the administration process the goal is to generate the exact same scenario stimuli for each trainee being assessed worldwide – the only difference being *when* the stimuli would be generated based on individual progress through the assessment. Such ideal consistency approaches feasibility when the assessment stimulus generation is standardized and automated. Being intensely systematic about maximizing reliability would give decision makers and trainees great confidence that the assessment is fair.

Maximize Training Utility

Assessment design and training design should be approached in an integrated fashion rather than separately (Pritchard, DiazGranados, Weaver, Bedwell, & Harrell, 2009). A standardized assessment can retrospectively provide valid feedback for training and learning processes that have already occurred and prospectively drive future instructional prescriptions as well (Freeman et al., 2009). As Table 2 (adapted from Hogarth, 2001) shows, since these tasks have great real world and administrative consequences – lives may depend on their proper execution and may count toward promotion points – pressure for optimizing instructional strategies (what instructors can do to increase the probability of learner success) and learning strategies (what learners can do to increase the probability of learner success) will exist and optimal strategies can emerge.

Table 2. Feedback and Strategies

		Quality of Feedback	
		Perfect	**Noisy**
Consequences of Errors	**Large**	Pressure for valid strategies – good strategies can emerge	Difficult for valid strategies to emerge
	Small	Little pressure for valid strategies – good and bad strategies can coexist	Superstitious learning quite likely

To find the best strategies, against what criteria measures should these strategy decisions be made? In addition to the normal point system used to measure effectiveness, because speed of execution is an important factor in saving lives on the battlefield, it will be important to calculate time-efficiency (e.g., points per second) as a second-order measure of the current level of expertise (Freeman et al., 2009). Take the case where two trainees obtain a perfect score on a task for effectiveness, but Trainee B takes half the time of Trainee A. Who is the better performer? Clearly, Trainee B is better and should be recognized and rewarded as such with a higher score. The presence of automation and recording of the assessment makes the tracking of time and the calculation of efficiency easy to do.

How perishable is the training the trainees receive? Certain instructional strategies are better for retaining these skills than others. The impact of these strategies can be determined by calculating a retention measure. For example, calculating the ratio of the new pretest score with old posttest score (assuming assessment equivalence) could provide such insight. A forgetting curve could be calculated to inform training frequency decisions.

How emotionally appealing was the training and how might a trainee's feelings change as a result of the training? A standardized satisfaction survey administered after the training would represent a trainee's subjective evaluation of the training experience and could help improve assessment and training processes. Also, affective measures could be taken after the performance pretest and then again after the performance posttest to determine affective gain. For example, engaging in stress exposure training after the pretest might well result in decreasing anxiety and increasing performance confidence for the Soldier (Driskell & Salas, 2009).

A trainee's ability to minimize the effects of battlefield trauma depends upon the three dimensions of cognitive, skill, and affective performance. For skill performance measurement, subjectively (Milham et al., 2009) or objectively weighted effectiveness and efficiency metrics could be combined into a multiattribute utility score to characterize the current skill level for an individual Soldier. Likewise, a different multiattribute utility measure with appropriate weightings could be calculated to combine the effectiveness, efficiency, retention, and appeal measures into one omnibus score for training purposes (Bewley et al., 2009). Maximizing this measure could then serve as the objective criterion for making rational training decisions.

Minimize Administration Demands

The design should seek to minimize the time and cost required to administer the assessment. Automation can help in this regard. For example, automated mannequins with biophysics-based reactions to treatment can replace soldiers acting as human casualties.

Minimize Maintenance Demands

The design should seek to minimize the time and cost required to maintain the assessment through its development cycles.

CONSTRUCTION

Develop an Administration Process

In order to be standardized, flexible, and avoid fatigue confounds, the assessment process will consist of the Army Core Skills assessment followed by the optional Unit Skills assessment. Core Skills will always be conducted inside to avoid weather confounds. The trainee will be alone in a mixed reality validation room to avoid any evaluator influence and social loafing from other trainees. There will be standardized recorded administrative messages, standardized scenarios including number of casualties, standardized initial simulator settings including number, location and severity of wounds, standardized reaction of simulated casualties, standardized scenario stimuli such as combat noise, and the trainees carrying a standardized First Aid Kit. Since this refresher training will occur yearly, these standardized scenarios should be changed yearly as well to prevent "test rehearsal".

The optional Unit Skills are conducted in a likewise systematic manner, whether indoors or outdoors.

Develop Performance Tasks

A performance task consists of an Army-assigned number (not yet assigned to CLS tasks), an Army-assigned task name, and the specified performance steps required to accomplish the task. The goal is to have a standardized scenario sequence, standardized tasks within the scenario, and standardized discrete observable steps for each standardized task within each scenario.

Task Inclusion

Which tasks will be assessed? Existing documentation of the informal assessment processes being used – often incomplete, contradictory, or incorrect – was used to build the task and step lists. In no case was a task deleted or a step ignored, although it may have been restated. In no case was a new performance task or step generated, as this was beyond the scope of the effort.

Currently, the Combat Lifesaver assessment consists of 9 tasks and more than 100 steps. The Medic assessment consists of 58 tasks and more than 1700 steps.

Step Clarity

In order to increase assessment reliability and objectivity, a well-stated performance step will "describe observable and measurable behaviors that reflect the most essential characteristics of good performance on a task, are clearly and briefly stated, and are written in language that students will understand" (Mueller, 2005, p. 6). In other words, a step should be relevant, demonstrable, and unambiguous. Relevance will be determined by appropriate medical authorities and is beyond the scope of this project; therefore existing official documentation was used to establish *de facto* relevance and medical correctness for the Core Skills Tasks. Demonstrable means the step is observable and is stated using an action verb in the past tense. Unambiguous usually means a single distinct physical action capable of using the binary Go/No-Go scoring scheme favored by the Army. Ideally, the single required behavior in the step will either be observed or it will not. This clarity should result in greater trainee satisfaction and better performance (Mueller, 2005).

If the task step as stated in the documentation is not demonstrable and unambiguous, several options exist (in order of desirability): restate the essence of the step, or; have trainee verbalize the essence of the step, or; accept the flawed step as is, or; eliminate the step.

As an example of restatement, the task step "Took/verbalized body substance isolation (BSI) precautions. (Go/No-Go)" is evaluated repeatedly in the Medic assessment. *TC 8-800* (p. 67) states that BSI means wearing, as a minimum, gloves and eye protection. However, one evaluator might award a Go to Trainee A for just wearing/verbalizing the gloves (the case of evaluators at one MSTC – none knew of the eye protection requirement), whereas a different evaluator might require just eye protection to award a Go to Trainee A, while a third evaluator watching Trainee A might require both in order to pass. In each case the evaluator honestly thought Trainee A "Took/verbalized body substance isolation (BSI) precautions." However, the assumptions varied between evaluators. The solution to both evaluator ignorance and elimination of evaluation ambiguity is to break up one ambiguous step into two unambiguous steps: "Wore/verbalized gloves. (Go/No-Go)" and "Wore/verbalized eye protection. (Go/No-Go)". An additional benefit of making tasks discrete is the increased likelihood that performance measurement processes can be captured automatically (Salas & Rosen, 2009) – an explicit requirement – which will decrease costs (Freeman et al., 2009).

Alternatively, the step can be restated to enable verbalization of the task. Consider the step "Evacuated the casualty. (Go/No-Go)". Actually performing this task as stated would be time-intensive, physically difficult for a one-room assessment using sensitive automated mannequins with no vehicles and no human

assistants, and of limited medical performance diagnostic value in the first place. The step can be restated as "Verbalized the need to evacuate the casualty" without losing the key medical insight that the casualty needed to be evacuated. Likewise, the step "Evaluated pulse, motor, sensory (PMS)" is found in numerous tasks and sometimes several times within the same task. After demonstrating the ability to perform the step the first time, the step can be restated as "Verbalized the need to evaluate pulse, motor, sensory (PMS)" without losing the key medical insight that the casualty's PMS needed to be evaluated at this point in the task. Verbalizing can save valuable assessment time.

Verbalizing some task steps may also provide insight into training problems (e.g., misconceptions) lurking below the Go/No-Go construct. Although this method has some limitations (e.g., some trainees may find it difficult to verbalize), verbal protocols "can provide vital information about the learner's cognitive processes, beyond simple measures of accuracy and time on task" (Trickett & Trafton, 2009, p.333). For example, analyzing verbalized mistakes can provide valuable insights to instructors for repairing flawed mental models (Klein & Baxter, 2009).

Develop a Scoring Rubric

To score effectiveness, evaluator step scoring is binary (Go/No-Go) for each step. This one or zero is then multiplied by the repetition weighting (e.g. a 5-2-1 scheme for the three possible attempts) to incentivize assessment preparation by the Soldier and minimize remediation costs. The sum of the step scores is then multiplied by a task weighting and a task-individualized scoring scheme (e.g., failing a designated critical step automatically fails the entire task) to obtain the task score.

For example, the Core Skills task weight is the importance of this task compared to other tasks as determined by subjective or objective criteria (e.g., by the U.S. Army Medical Department Center and School (AMEDD)) and reassessed as appropriate. Reflecting current operational realities, the weighting scheme counts hemorrhage control tasks as a 9 (since 9% of preventable battlefield deaths are of this type), 5 for needle chest decompression (same reason), 2 for airway management (same reason) and 1 for all other tasks. This ensures the scoring is biased toward tasks having the greatest impact on minimizing preventable battlefield fatalities.

The sum of the weighted task scores is the total score for the Core Skills. Optionally, the Core Skills could be added to the Unit Skills score to determine the individual's grand total score.

Importantly, evaluators will be able to manually override the standardized scoring system for tasks and overall to allow for unconventional-but-successful end state solutions (e.g., the casualty survives). The required evaluator justifications in these rare cases will be fertile ground for assessment system improvements.

Army policy is that external sources should evaluate training whenever possible to objectively measure performance in terms of Army and joint standards. In this case, having a cadre of external impartial professional evaluators (e.g., from AMEDD in San Antonio, Texas) using secure MSTC webcams anywhere in the world could minimize unit bias in scoring. This would enable valid Army-wide competence assessment for the Army Core Skills. Alternatively, random checks on the primary MSTC evaluators by these distant evaluators (e.g., a 10% sampling), may also help minimize any local unintentional bias. This kind of distant evaluation through a video stream may be unique in Army training.

Minimize Administration Demands

The administration process will minimize the time required by: (1) right sizing the content to be assessed (e.g., all tasks trained in CLS/Medic curricula and no more); (2) paring repetitive tasks (e.g., demonstrate trauma casualty assessment capabilities only once – thereafter verbalize the need to assess when it is part of another task), and; (3) if MSTCs use the now-enabled "smart training" customized approach (including data from the cognitive assessment), only tasks failed during the single-attempt pretest will be trained. After this remediation training (Milham et al., 2009), which may occur over several days through instructor and automated means, the trainee will take the posttest. If a scenario task is failed, the trainee will be able to retrain immediately (using the customized training plan) after each failed posttest task, get back into line and retake the failed scenario two more times.

The administration process will minimize the cost required by: (1) leveraging existing/planned resources (e.g., documentation, MSTC automation); (2) minimizing the human resources required (e.g., as live casualties and evaluators), and; (3) seeking to minimize the use of real-world consumables (e.g. needles, bandages, etc.).

Minimize Maintenance Demands

The support system surrounding the assessments will be centralized, streamlined, and automated to the extent possible. For example, there should be a single point of responsibility (e.g., AMEDD Department of Combat Medic Training) for developing and maintaining the assessments Army-wide. A sensor infrastructure should be installed (e.g., automated medical records to determine task weighting, appropriate task steps from the field and the medical literature, etc.). An analytical infrastructure should be installed (scoring analysis, content validity ratio determination, trends, automated training recommendations, etc.). Finally, an information dissemination infrastructure should be installed (e.g., email lists).

VERIFICATION AND VALIDATION

The proposed medical tasks, steps and scoring scheme will be validated through official documentation, SME interviews, and SME email surveys. The proposed administration process and technology will be verified through a proof-of-concept demonstration (Stanney, 2009) at a MSTC.

Validation: Subject Matter Experts

To obtain initial feedback, interviews of several senior NCO medics and MSTC personnel were conducted. Reaction to the proposed assessments was generally positive. For example, the proposed assessment was characterized by one senior NCO medic as a "potentially excellent tool" in that it would: (1) provide schoolhouse and operational commanders insight into Army-wide first responder issues and trends which it currently does not have, resulting in better situational awareness; (2) result in better training, and; (3) increase the confidence of medics and combat lifesavers.

Scoring feedback has been incorporated into the assessment by: (1) including the possibility of random, rather than continuous, scoring by AMEDD cadre to ensure MSTC cadre scoring objectivity; (2) attempting to follow the principles of outcome based training and education (OBT&E), specifically by allowing manual task score and overall score overrides – to get to the desired end state – in innovative ways; (3) confirming the value of the 5-2-1 weighting system for step attempts, and; (4) confirming the concept of 70% of the total score available as the passing criterion.

Verification: Demonstration

A successful content, process, and technology demonstration will confirm the feasibility of these standardized assessments.

To this end, a software evaluator tool (Figure 1) will be developed for table or tablet PCs to facilitate assessment planning, execution, and data analysis. Functions include: startup (e.g., user authentication);

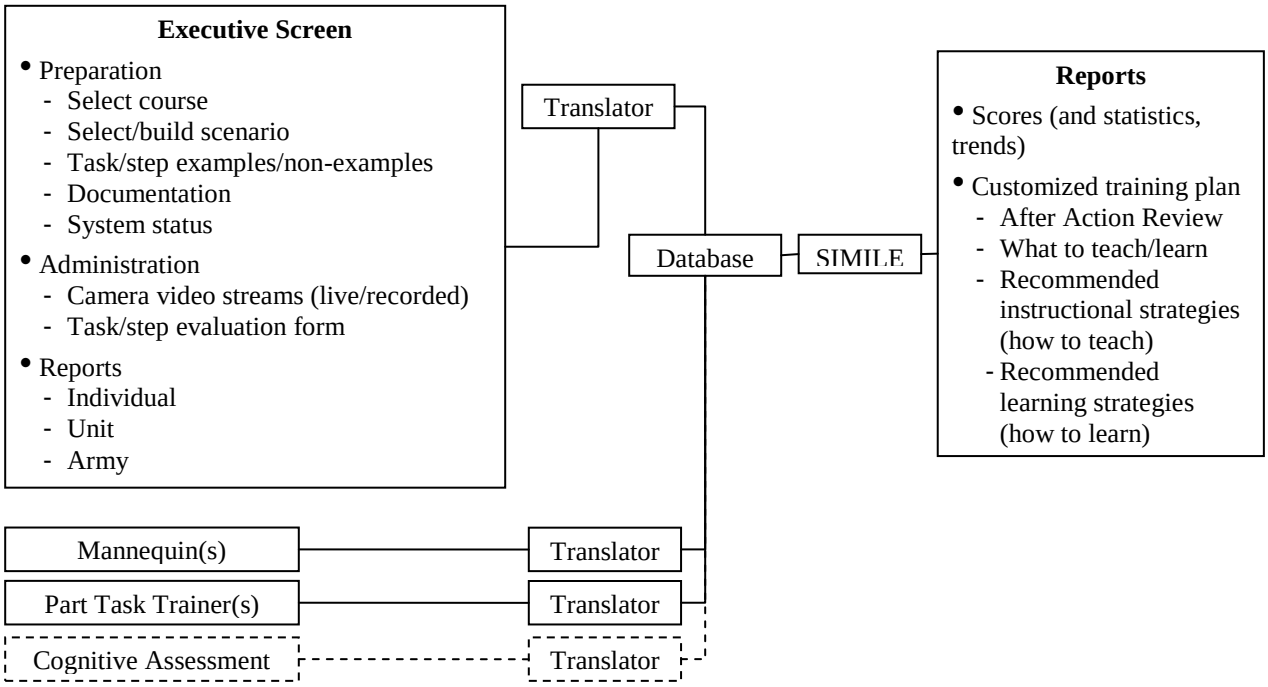

Figure 1. Simplified demonstration functional architecture

orientation (e.g., system overview); preparation (e.g., scenario development, scoring rubric setup (e.g., unit task weights), process checklists (e.g., supplies, evaluator training, task/step review (e.g., example/non-example videos); assessment execution (e.g., evaluation score sheets), post assessment (e.g., customized training plan, after action review (Milham et al., 2009), trends, archives), and supporting activities (e.g., search, print, database activities).

For example, XML-based Simulations for Integrated Learning Environments (SIMILE™) software will integrate performance data automatically collected from simulation technologies (e.g., mannequins and part-task trainers), evaluator task performance scoring, standardized and flexible scoring rubrics, and reporting requirements.

CONCLUSIONS

As a result of meeting requirements and design goals, this prototyping effort will help the Army move from a system of informal non-standardized assessments to formal standardized assessments for battlefield first responders and provide a firm foundation for future development. The result will be more scientifically defensible, more comprehensive, more automated, and more consequential assessments. Most importantly, these MSTC-based assessments of first responder skills will be more useful, as they will inform training and operational decisions that could minimize preventable deaths and trauma on the battlefield.

ACKNOWLEDGEMENTS

The authors wish to express their gratitude to our funding agency, U.S. Army Program Executive Officer for Simulation, Training, and Instrumentation – Assistant Program Manager MedSim of Program Manager Combined Arms Tactical Trainers, and to others who have contributed to the project: Bill Pike, Chris DuBuc, Dawn Riddle, and Roger Chapman.

REFERENCES

Asymmetric Warfare Group (2009). Outcomes Based Training and Education (OBT&E): An Introduction to the Idea. Retrieved June 8, 2009, from https://atn.army.mil/Media/docs/0120_OBTE_UNK.pdf

Bewley, W.L., Chung, G.K.W.K., Delacruz, G.C., & Baker, E.L. (2009). Assessment models and tools for virtual environment training. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond* (pp. 300-313). Westport, CT: Praeger Security International.

Carmines, E.G. & Zeller, R.A. (1979). *Reliability and validity assessment*. Newbury Park, CA: Sage Publications.

Driskell, J. & Salas, E. (2009). Affective measurement of performance. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond* (pp. 362-375). Westport, CT: Praeger Security International.

Freeman, J., Stacey, W., & Olivares, O. (2009). Measurement and assessment for training in virtual environments. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond* (pp. 236-250). Westport, CT: Praeger Security International.

Hogarth, R.M, (2001). Cited in Goldstein, D.G., Hogarth, R.M., Kacelnik, K., Kareev, Y., Klein, G., Martignon, L., Payne, J.W., & Schlag, K.H. Group report: Why and when do simple heuristics work? In G. Gigerenzer & R. Selten (Eds.), *Bounded rationality: The adaptive toolbox*. Cambridge, MA: The MIT Press.

Klein, G., & Baxter, H. (2009). Cognitive transformation theory: Contrasting cognitive and behavioral learning. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond* (pp. 50-65). Westport, CT: Praeger Security International.

Messick, S. (2000). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. In R. Goffin & E. Helmes (Eds), *Problems and solutions in human assessment:*

honoring Douglas N. Jackson at 70 (pp. 3-20). Norwell, MA: Kluwer Academic Publishers.

Milham, L., Carroll, M.B., Stanney, K., & Becker, W. (2009). Training systems requirements analysis. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond* (pp. 165-192). Westport, CT: Praeger Security International.

Mueller, J. (2005). The authentic assessment toolbox: enhancing student learning through online faculty development. *Journal of Online Learning and Teaching, 1*(1), July, 2005.

Pritchard, R.D., DiazGranados, D., Weaver, S.J., Bedwell, W.L., & Harrell, M.W. (2009). Virtual environment performance assessment: Organizational level considerations. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond* (pp. 284-313). Westport, CT: Praeger Security International.

Salas, E. & Rosen, M.A. (2009). Performance assessment: Section perspective. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond* (pp. 227-235). Westport, CT: Praeger Security International.

Stanney, K. (2009). Requirements analysis: Section perspective. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond* (pp. 115-130). Westport, CT: Praeger Security International.

Trickett, S., & Trafton, J.G. (2009). A primer on verbal protocol analysis. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond* (pp. 332-346). Westport, CT: Praeger Security International.

Trochim, W. M. (2009). Measurement validity types. In *The Research Methods Knowledge Base, 2nd Edition*. Retrieved June 8, 2009, from http://www.socialresearchmethods.net/kb/measval.php