Evaluation and Outcomes of a Novel AAR Scoring Methodology

Taryn Cuper, Phillip Jones, David Drucker
MYMIC, LLC
Portsmouth, VA
Taryn.Cuper@mymic.net,
Phillip.Jones@mymic.net,
David.Drucker@mymic.net

John Morey, PhD

Dynamics Research Corporation
Andover, MA

JMorey@drc.com

ABSTRACT

This paper describes an after-action review (AAR) behavior scoring system developed for a study of a medical simulation center's support to National Guard small team clinical training. The application of this AAR scoring system provides empirical insight into the behaviors key to AAR success. A significant part of modern training, the AAR is a professional, facilitated discussion within a training audience that compares trainee performance against task standards and training objectives. A great deal of study and theory has been developed on the conduct of an AAR. However, there are no quantifiable measures for enumerating the effectiveness of an AAR and AAR behaviors. Such measures would be valuable in determining the pertinent aspects of the AAR that facilitate the most effective experience for the training audience. The methodology presented in this paper, based on established AAR learning theory, describes fifteen AAR facilitator behaviors organized across four AAR elements as well as five training audience AAR behaviors and five overall AAR measures of performance. Two analysts scored thirty-two separate, small team AARs across these twenty-five measures. The scoring included both a simple count of behaviors and a behavior quality score. The scoring yielded an inter-rater agreement (Kappa) of 0.53, which is a moderate strength of agreement, instrumental in validating the scoring process. The paper covers the development, execution, and analytical results of the scoring method. The results indicate the specific behaviors most correlated to overall AAR success as well as contradicting some conventional wisdom regarding the AAR process. Implications of these findings on future AAR evaluation research are addressed.

ABOUT THE AUTHORS

Taryn Cuper is the Deputy Director for MYMIC's Analysis, Training, and Assessment Group, with a focus on assessing the application of modeling and simulation solutions against organizational goals. She serves as MYMIC's lead for medical domain solutions and for research standards development and execution. Ms Cuper has a background in cognitive science and modeling and simulation.

Dr. John Morey is a board certified human factors professional with 30 years experience in applied training and human factors research and development. For the past 13 years he has been conducting team research in health care, specializing in team performance measurement.

Mr. Phillip Jones is the Director of MYMIC's Analysis, Training and Assessment Group. He has over 22 years of professional training experience as an Army combat-arms officer, serving as Observer/Controller at an Army Combat Training Center and instructing AAR facilitators.

D. Nick Drucker is a research analyst in MYMIC's Analysis, Training and Assessment group. Mr. Drucker has a background in International Relations and Psychology and is currently pursuing an MA in International Relations, focusing on modeling and simulation within the field.

Evaluation and outcomes of a novel AAR scoring methodology

Taryn Cuper, Phillip Jones,
David Drucker
MYMIC, LLC
Portsmouth, VA
Taryn.Cuper@mymic.net,
Phillip.Jones@mymic.net,
David.Drucker@mymic.net

John Morey, PhD

Dynamics Research Corporation
Andover, MA

JMorey@drc.com

INTRODUCTION

The United States Army developed the After Action Review (AAR) over time as a process to support collective training. The definition of an AAR found in Training Circular (TC) 25-20 (1993) is:

> An after action review (AAR) is a professional discussion of an event, focused on performance standards, that enables soldiers to discover for themselves what happened, why it happened, and how to sustain strengths and improve on weaknesses. It is a tool leaders and units can use to get maximum benefit from every mission or task. It provides
>
> - Candid insights into specific soldier, leader, and unit strengths and weaknesses from various perspectives
> - Feedback and insight critical to battle-focused training
> - Details often lacking in evaluation reports alone

Typical small unit leaders have stated that AARs are critical to extremely critical to training success, and that 25-35% of the total training benefit occurred in the AAR (Jones & Mastaglio, 2006).

Army AAR methodology contained in doctrinal publications describes how to conduct an AAR, however, it does not specify the behaviors required to carry out this guidance, i.e. the process knowledge required to becomes an expert (Jones & Drucker, 2009). There has been an effort to study the AAR process to determine its components and capabilities. One such effort was done by Morrison and Meliza (1999) who identified theories and techniques for successful AARs. The Army is conducting efforts to better understand the theoretical basis of the AAR and to identify the knowledge, skills, and abilities (KSA) associated with participating in an AAR, either as the facilitator or the training audience. Currently, with some exceptions these KSA exist as tacit knowledge among AAR practitioners, learned primarily via apprentice-based learning (Jones & Mastaglio, 2006). Referenced exceptions can be found in local materials such as standard operating procedures and AAR "rules of engagement" (O/C Handbook, 2006). This paper documents a novel approach to analyzing the AAR by implementing a scoring method, focused on facilitator (FAC) and training audience (TA) behaviors, that may allow the objective rating of AAR effectiveness.

BACKGROUND

Applying the AAR Scoring Method under Experimental Conditions

The scoring methodology was developed in support of a study on the effectiveness of using civilian medical simulation centers to support training of military medics. This included the ability of the center to support AARs. The training consisted of groups of medics treating "injured" human patient simulators (HPS). There were a total of seventy-four trainees from both the Army ($n = 46$) and the Air Force ($n = 28$). Trainees were divided into training groups ranging from three to five personnel. Each group consisted entirely of Army or Air Force personnel; no groups had a combination of the two services. Trainee ranks ranged from Private (E2) through senior Sergeants with the most frequent rank being E4. There were a total of nine facilitators. Facilitator ranks ranged from junior NCOs (E5) through senior officers (O6), with the majority of facilitators being Sergeants (E5).

The HPSs, with appropriate moulage, provided realistic physiological stimulus to the training audience. This was augmented by simulation managers role-playing the patients aurally from a control room and additional scenario information provided by the facilitator. Each drill consisted of one patient. Site management, to include resource management, was not a part of the training.

A training session consisted of four training drills, each drill followed by a corresponding, non-cumulative AAR. There was no situational continuity between the four drills. Within each training session, a single individual facilitated each of the four AARs while a second individual served as an observer/evaluator. For this portion of the study, only the actions of the primary facilitator were considered. To provide some degree of consistency, each facilitator was given standardized instructions prior to facilitation. Two facilitators led more than one session, one conducting three sessions and another conducting two. The experiment consisted of sixteen training sessions, resulting in sixty-four AARs, thirty-two standard and thirty-two augmented. However, there was no record for one AAR, resulting in its exclusion from analysis and a total AAR count of sixty-three. Each AAR was recorded and categorized by day, group, and scenario.

As part of the experimental design, AARs were conducted under two situations: standard and augmented. Standard AARs used no extraneous materials to aid in AAR facilitation, other than a blank flipchart. Augmented AARs included a flip chart with pre-determined AAR structure for the facilitator to follow. Eight training sessions conducted standard AARs and eight conducted augmented AARs. Time was available to the training audience following each AAR for them to internally discuss performance (i.e. further AAR) and adjust/plan for the next drill.

METHODOLOGY

Developing the Scoring Methodology

In order to create a method for analyzing AAR effectiveness, the research team created a unique AAR behavior scoring methodology. The foundation for this effort was a review of a collection of AAR sources that together described the foundations, development, and practice of AAR. These include an Army training circular, which provided instruction on planning and preparing for, as well as conducting, the AAR. This also included two reports from the Army Research Institute covering AAR methods, practices, and products developed from behavioral science principles, and effectiveness analyses of AARs conducted at the Combat Training Center (Salter & Klein, 2007; Morrison & Meliza, 1999). All sources contributed to the identification of accepted and effective AAR practices on the part of facilitators, expected benefits of the AAR for the training audience, and recommendations for how best to realize an optimal AAR experience.

Behaviors were selected based upon their perceived importance to overall AAR success. Emphasis was placed on facilitator behaviors (17 scored behaviors) over training audience behaviors (five scored behaviors). Facilitator behaviors were grouped into the following categories:

- General protocol: i.e., asking open-ended questions, positive reinforcement of performance, etc.
- Grounding: relating training experience to ground truth, actual performance, and operational conditions
- Structure: orientation to training objectives, and group performance
- Information transfer: mentoring, referencing metrics or other published information

All behaviors were scored in two ways, by count and by quality. Raters indicated how many times they observed a given behavior as well as rated the quality on a Likert scale of 1 (Poor) to 5 (Excellent), as indicated in Table 1. Ten general AAR quality scoring criteria were added, from which an overall AAR quality score was calculated. These also were rated on Likert scale of 1 (poor) to 5 (excellent); an AAR could therefore have a maximum overall score of 50. Table 1 provides the descriptors for each level. Tables 2-4 list the facilitator behaviors, training audience behaviors, and general AAR quality criteria below.

Table 1. AAR Behavior Quality Scale

Score	Descriptor
1. Poor	Performs behavior without verbal or other reinforcement

	(e.g. only gestures to encourage participation)
2. Fair	Performs behavior with limited reinforcement or detail (e.g. asks, "What did you think?" without further clarification or question context)
3. Good	Performs behavior with adequate reinforcement and detail (e.g. asks, "What do you think went well?")
4.Very Good	Performs behavior with additional detail (e.g. asks, "John, what did you think about how the team addressed the airway?"
5. Excellent	Performs behavior with a great deal of detail and reinforcement (e.g. "asks, John, how did the team respond when you found the patient could not be intubated?)

Table 2. Facilitator Behaviors

Asks open ended question
Asks yes / no question
Encourages member to participate
Prevents or corrects attribution/blame
Lays blame on an individual
Positive reinforcement of performance
Positive reinforcement of participation
Assists in determining actual conditions /GT
Assists in determining actual performance
Emphasizes operational vs exercise conditions
Orients AAR to training objectives
Orients AAR to group vs individual performance
Puts TA/AAR back on track
Coaches or mentors group
Projects changes to future performance
Refers to specific published metrics/ knowledge
Disperses/injects knowledge

Table 3. Training Audience Behaviors

Disperses/injects knowledge
Self-Identifies individual sustain or improve
Self-Identifies group sustain or improve
Provides rationale for performance
Projects changes to future performance

Table 4. AAR Quality Criteria

TA identifies and accepts what actually occurred
TA identifies what to sustain and what to improve
TA projects changes into future
FAC identifies what actually occurred
FAC identifies what to sustain and what to improve
FAC projects changes into future
Level of TA participation
Level of FAC participation
AAR is focused on performance
AAR concludes on a positive note

Analysis of Data

AAR scoring is a novel practice. As such, for this effort, analysis was treated as an exploration of the data with a focus on determining the feasibility or usefulness of such an evaluation. This exploration was conducted along two interactive lines of inquiry. The investigators first conducted a statistically informed, quantitative evaluation. The results of this evaluation were then reviewed by an AAR subject matter expert (SME) for qualitative evaluation and for the identification of further statistical inquiries.

Two independent raters scored each AAR, with the criteria that if a Cohen's Kappa measure of inter-rater agreement was unacceptable, the raters would regroup to further refine the methodology and then rescore. While it was not possible to determine inter-rater agreement for behavior frequency and quality due to the manner of scoring, Cohen's Kappa was used to determine the level of agreement between raters with regard to whether a behavior was at all present. This measure serves as a first step toward verifying that the chosen scoring behaviors are indeed distinct and identifiable. Cohen's Kappa is a measure of inter-rater

reliability, considered more robust than a percent agreement calculation because it takes into account agreements that occur by chance. The formula from which Kappa is calculated is as shown in equation 1:

$$\frac{\text{(Percent Agreement - Percent Chance Agreement)}}{\text{(1 - Percent Chance Agreement)}} \quad (1)$$

FINDINGS

General Findings

The AAR scoring methodology, which started as an experiment within the experiment, proved to be more successful than originally anticipated. The methodology showed that AARs could be scored for facilitator and training audience behaviors and that this scoring could be used for varied analysis to understand and evaluate the dynamics occurring within the AAR.

Ability to Score AARs

Using the methodology, two observers were able to relatively easily score the sixty-three AARs. On average, the AARs lasted just over ten minutes. Scoring took approximately seventeen minutes per AAR, or somewhat longer than AAR duration.

The scoring process was conducive to reporting nearly every behavior observed. The total count of behaviors averaged between the two observers was 1,629, or approximately twenty-six behaviors per AAR. Inter-rater agreement indicated that the chosen scoring behaviors were valid for further inquiry and analysis. A Kappa between 0.40 and 0.60 is considered a moderate level of agreement. The Kappa level for the raters in this effort was 0.53, well above the predetermined threshold for rescoring.

The observers were also able to use the AAR Overall Quality scoring to set a relative AAR assessment. On average, the AARs scored above satisfactory (see the distribution of scores in Figure 1). Three facilitator behavior rating items, *Level of Facilitator Participation*, *AAR Focused on Performance*, and *Facilitator Focused on Performance*, had the highest mean ratings, 4.36, 4.22, and 4.03 respectively (on a 1 to 5 scale where 5 is excellent). Two training audience behavior rating items, *Training Audience Projects Changes into the Future* and *Facilitator Projects Changes into the Future* were rated lowest, at 1.4.

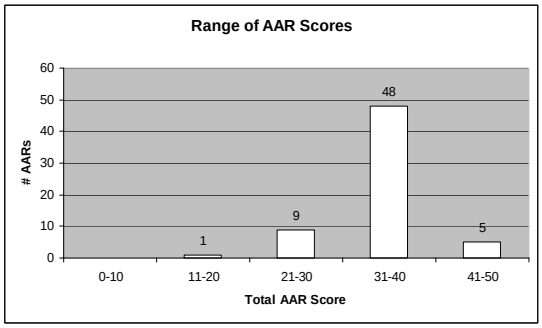

Figure 1. Range of AAR Scores

Use of Data to Analyze AARs

Descriptive Statistics

For facilitators, the behaviors of *Asking Open Ended Questions* and *Coaches or Mentor Groups* were the most frequently observed, with average counts of 6.40 (SD = 3.84) and 3.73 (SD = 1.91) observations per AAR, respectively. For the training audience, the behavior of *Self-Identifies Individual Sustains or Improves* was the most observed, with an average count of 2.51 (SD = 1.3) per AAR. The least observed behaviors included the *Facilitator Putting the AAR Back on Track* (*M* = 0, SD = 0) and *Preventing or Correcting Attribution or Blame* (M = 0.04, SD = 0.16), and Refers to Published Metrics (M = 0.10, SD = 0.22), and the *Training Audience Lays Blame* on an Individual (M = 0.05, SD = 0.17). Two of these behaviors, *Puts AAR Back on Track* and *Prevents Attribution* necessitate situational precursors, i.e. in order to put the AAR back on track, the AAR would have to be going off track. Thus, these behaviors are dependent upon the actions of the training audience within the AAR. The relative short duration of the AAR and the small size and professional nature of the group might have served to prevent behavior precursors.

From the perspective of quality, the facilitator behaviors *Coaches or Mentors Group*, *Positive Reinforcement Of Performance*, and *Asks Open Ended Questions* were all scored as higher than satisfactory with averages of 3.77 (SD = 0.79), 3.31 (SD = 0.82), and 3.24 (SD = 0.39) respectively. *Self Identifying Sustain or Improve*

was scored higher than satisfactory for the training audience, with an average of 3.40 (SD = 0.85). The quality of *Projecting Changes to Future Performance* was scored low for both facilitators and training audiences.

Regression Analysis

Because this study incorporated a new method to examine AARs, we wanted to scrutinize the data to explore for possible relationships. We used regression analysis to look for these potential relationships and used the results to aim our qualitative analysis. We looked for relationships between: facilitator behaviors, and overall score, training audience behaviors and overall score, and between the facilitator and training audience behaviors. As a result, our independent and dependent variables are a mix of training audience behaviors, facilitator behaviors and the overall AAR score. All variables in these models were behavior scores. No significant results were found in counts. The bi-variate regressions involved the facilitator behavior *Emphasized Operational vs. Exercise Conditions*, and *Overall AAR score* as the dependent variables, with *Training Audience projects changes to future performance*, and *Training Audience Disperses/Injects Knowledge* as the independent. Four multi-variate regression analyses were also performed. The dependent variables for these models were: *Training Audience Identifies and Accepts What Actually Occurred*, *Overall Score on AAR* (used for two), and *Level of Training Audience Participation*. For the first model, the independent variables were the *Level of Facilitator Participation* and *Facilitator Identifies/Accepts What Actually Occurred*. The second model's independent variables were *Facilitator Asks Open Ended Questions* and *Facilitator Coaches or Mentors Group*. The independent variables for the third model were *Facilitator Encourages Member to Participate*, *Facilitator Asks Open Ended Questions*, and *Positive Reinforcement of Participation*. The final model's independent variables were *Facilitator Asks Open Ended Questions*, *Facilitator Encourages Member to Participate*, *Facilitator Assists in Determining Actual Performance*, and *Facilitator Assists in Determining Actual Conditions/Ground Truths*.

Regression analysis showed a positive relationship (p = 0.012, R^2 = 0.099, regression coefficient = 0.496) between the quality of *Facilitator Emphasizes Operational vs. Exercise Conditions* and the quality of *Training Audience Projects Changes to Future Performance*. This result corresponds to a tenant of adult education that adults learn – or change behavior – only if they see a personal benefit. It follows that a training modality must have a level of acceptance by the training audience; the audience must feel that the results of and observations from training are legitimate to their operational requirements. Thus, it is critical that the facilitator, who is seen as an authority figure within the training paradigm, supports the legitimacy of the training modality by relating it to actual operational conditions. Supporting this observation is the positive relationship (p = 0.004, R^2 = 0.135, regression coefficient = 0.630) between the quality of *Facilitator Identifies and Accepts What Actually Occurred* and the quality of the *Training Audience Identifies and Accepts What Actually Occurred*. The facilitator identifying and accepting a "ground truth" encourages the training audience to identify and accept a "ground truth".

Similarly, there is a positive relationship (p = 0.017, R^2 = 0.090, regression coefficient = 2.106) between the overall score of the AAR and the quality of *Training Audience Disperses / Injects Knowledge*. It is interesting that the relationship is between the quality of the AAR and the training audience injecting knowledge versus the facilitator injecting knowledge. This demonstrates that the quality of the AAR belongs to and is dependent upon the training audience, not the facilitator.

The results also illustrate the importance of asking open-ended questions to facilitate an AAR. Open-ended questions are the best method of eliciting audience participation, audience initiative, and of getting the audience to, in turn, elicit and socialize the tacit knowledge developed as part of the training experience. This is reflected in the following:

- There is a positive relationship (p = 0.000, R^2 = 0.311, regression coefficient = 13.306) between the overall score of the AAR and the quality of *Facilitator Asks* Open-*Ended Questions*.
- There is a positive relationship (p = 0.007, R^2 = 0.138, regression coefficients = 1.486) between the quality of *Facilitator Asks Open-Ended Questions* and training audience participation.

- In addition, other ways of improving training audience participation can also improve the quality of an AAR. The positive relationship (p = 0.023, R^2 = 0.368, regression coefficients = 2.261) between the overall score of the AAR and the quality of the *Facilitator Encourages TA Participation* demonstrates this finding.

Exploratory Independent T-Tests

In addition to the above analyses, multiple, independent t-tests were also performed to try and identify whether any relationships exist among the scored behaviors in our dataset. Multiple t-tests violate a classical assumption of the testing method by increasing the likelihood of Type I error. However, as this research is evaluating a novel scoring method and as the purpose of these tests was to look for possible relationships, the potential error was believed to be acceptable for analysis purposes. In future studies we would use more rigorous statistical testing, with more data, to reach conclusive results.

The independent t-test findings between standard and augmented AARs appear to indicate that augmentation tended to contribute towards a more formal AAR process. Augmented AARs showed significant increases in and higher quality of references by the facilitator to training objectives and published material. In turn, this increased structure and reference to goals might have facilitated training audience involvement, as the augmented AARs showed increased counts of the *Training Audience Dispersing/Injecting Knowledge*, *Self-Identifying Sustains and Improves*, and *Providing Rationale for Performance*. The following show findings of interest from behavior counts and quality scores, respectively.

Findings of interest from behavior counts:

- On average, *Facilitator Asks Open-Ended Questions* occurred **more** often in augmented AARS.
- On average, *Facilitator Prevents or Corrects Attribution/Blame* occurred **less** often in augmented AARs
- On average, *Facilitator Orients AAR to Training Objectives* occurred **more** often in augmented AARs
- On average, *Facilitator Refers to Specific Published Metrics*/*Knowledge* occurred **more** often in augmented AARs.
- On average, *Training Audience Group Disperse/ Inject Knowledge* occurred **more** often in augmented AARs.
- On average, *Training Audience Self-Identifies Group Sustain*/*Improve* occurred **more** often in augmented AARs.
- On average, *Training Audience Provides Rationale for Performance* occurred **more** often in augmented AARs.

Findings of interest from behavior scores:

- On average, the facilitators in augmented AARs scored 0.82 **higher** on *Orient AAR to Training Objectives*.
- On average, the facilitator in augmented AARs scored 0.33 **higher** on *Refers to Specific Published Metrics*/*Knowledge*.
- On average, the training audience in augmented AARs scored 0.73 **higher** on *Group Disperse*/*Inject Knowledge*.

Independent t-tests were also used to determine any significant differences in the overall AAR scores between both standard and augmented AARs and the top and bottom 20th percentiles. There was no difference between the overall quality ratings of standard versus augmented AARs. In the overall AARs quality ratings, a single training audience behavior, *Training Audience Identifies and Accepts What Actually Occurred*, was rated significantly higher on augmented versus standard AARs. However, this difference was found between average ratings of 3.53 (augmented) and 3.52 (standard); as such, the team concluded that such a finding is of no practical value and should not be considered further.

We compared participant behaviors in the top 20th percentile of overall AAR scores to the bottom 20th percentile to investigate the possible impact of facilitator and training audience behaviors on assessed AAR quality. Only the behaviors of the facilitator proved significant. The frequency (count) of the following behaviors was significant.

- 5.5 **more** instances of *Facilitator Asks Open Ended Questions* in the upper 20% of AARs.
- 2 **less** instances of *Facilitator Asks Yes/No Questions* in the upper 20% of AARs
- 1.5 **more** instances of *Facilitator Encourages Members to Participate* in the upper 20% of AARs
- 1.8 **more** instances of *Facilitator Positive Reinforcement of Performance* in the upper 20% of the AARs
- 1.5 **more** instances of *Facilitator Coaches or Mentors Group* in the upper 20% of the AARs

Based on average ratings, the following differences in quality were observed between the top and bottom 20th percentiles.

- Facilitators scored 0.5 points **higher** on *Asks Open-Ended Questions* in the upper 20% of AARs
- Facilitators scored 1.4 points **higher** on *Encourages Members to Participate*.
- Facilitators scored 0.66 points **higher** on *Coaches or Mentors Group* in the upper 20% of the AARs.
- Facilitators scored 0.73 points **higher** on *Refers to Specific Published Metrics/Knowledge* in the upper 20% of AARs.

The team also analyzed the effect of rapid, multiple AARs on facilitator and training audience behaviors, accomplished through t-tests* examining possible differences between behaviors from the first to the fourth AAR. Paired sample t-tests, grouping the behaviors by AAR number, were used and identified the behaviors that indicated considerable differences existed between the first and fourth AAR.

- Quality of *Training Audience Self-Identifies Individual Sustain / Improve* **improved**.
- Quality of *Training Audience Projecting Changes to Future Performance* **improved.**
- Quality of *Facilitator Asks Yes/No Questions* **improved.**
- Quality of *Facilitator Assists in Determining Actual Conditions / Ground Truth* **got worse.**
- Quality of *Facilitator Emphasis on Operational vs. Exercise Conditions* **improved.**

In comparing the first to the fourth AAR, significant development of performance enhancing skills was found. Both training audience behaviors *Self-Identifies Individual Sustain / Improve* and *Projecting Changes to Future Performance* improved. As participating in an AAR itself requires certain skills and abilities, this demonstrates that rapid, repeated AARs may serve to familiarize and train the training audience for AAR participation.

In contrast to the above, the quality of *Facilitator Assists in Determining Actual Conditions / Ground Truth* worsened. The interpretation of this result is dependent upon the definition of "assists." As the training audience becomes more accustomed to both the training modality and the AAR process, the training audience likely becomes self-facilitating and the facilitator might require a lower level of intervention or facilitation. Alternatively, this result may be indicative of the facilitators' level of fatigue with rapid, multiple AARs.

DISCUSSION

The AAR scoring methodology, though new and experimental, provided more success across a wider perspective than was originally expected. The scoring methodology provided several insights into the overall effectiveness of civilian medical simulation training environment. Two such insights included:

-Evidence of initial training participant dissonance caused by a new training environment and the dissipation of that dissonance in the improvement of AAR behaviors between the initial exposure to the environment in the first training drill and the performance in the fourth and last drill. The investigators assessed that this unique form of training, in order to maximize its effectiveness,

* The multiple t-tests used in this section suffer from the same violation as the previous set of tests. For the same reasons stated above we felt it was still necessary to employ these tests for analysis reasons.

requires a more extensive meta-training, or train-the-trainer effort.

-Low counts and quality of projecting current experiences into future operations, i.e. the lack of indicators that the participants were projecting this training event to their future organizational and operational requirements. Investigators postulated that this is due to issues with the participation abilities of facilitators and training audiences; projecting current behaviors into the future is an AAR skill. However, it could also be because of disconnects between the relatively alien nature of the civilian medical simulation center environment and the military operational medical environment or disconnects between the experiment's drill scenarios and the participants' mental models of operational scenarios. Significant disparities could prevent the training audience from modifying or replacing existing models.

There are several paths available for improving the scoring methodology. In the future, investigators would eliminate the negative behavior representations (*Fails to Correct Attribution* is the negative representation of *Corrects Attribution*) that did not conform to the Likert scoring approach. In addition, as the scoring methodology was developed and behaviors chosen, facilitators and training audiences were necessarily treated as discrete entities, with the bulk of attention placed on the facilitator. Execution of the scoring methodology, however, demonstrated a much more integrated and mutually dependent relationship between facilitators and training audiences, which needs to be taken into consideration going forward.

REFERENCES

Department of the Army. (2003). *Field Manual 7-1, Battle Focused Training.* Washington, DC: Author.

Department of the Army. (September 1993). Training Circular 25-20, A Leader's Guide to After-Action Reviews. Washington, DC: Author

Department of the Army, Joint Readiness Training Center (2006). *O/C Handbook*. Fort Polk, LA: Author

Jones, P. & Drucker, D (2009). [Review of Army AAR Literature, Army and Installation Levels], Unpublished raw data available from 140 University Blvd, Suite 100, Portsmouth VA 23703

Jones, P., & Mastaglio, T. (2006). *Evaluating the Contributions of Virtual Simulations to Combat Effectiveness* (Report Number 2006-04). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Kirkpatrick, Donald (1994) *Evaluating Training Programs: The Four Levels*. Berrett-Koehler Publishers, San Francisco

Morrison, J., & Meliza, L. (1999). *Foundations of the After Action Review Process*. (Special Report 42). Alexandria, VA: U. S. Army Research Institute for the Behavior and Social Sciences.

Salter, M. S. & Klein, G. E. (2007). *After Action Reviews: Current Observations and Recommendations*. (Research Report 1867). Alexandria, VA: U. S. Army Research Institute for the Behavior and Social Science