Development and Validation of Measures for Army Aviation Collective Training

Martin Bink

Melinda Seibert

Courtney Dean

John Stewart

Troy Zeidman

DEVELOPMENT AND VALIDATION OF MEASURES FOR ARMY AVIATION COLLECTIVE TRAINING

Martin Bink
U.S. Army Research Institute
Fort Benning, GA
Melinda Seibert
Courtney Dean
Aptima, Inc.
Woburn, MA
John Stewart
U.S. Army Research Institute
Fort Rucker, AL
Troy Zeidman
Imprimis, Inc.
Huntsville, AL

Simulation-based Aviation Training Exercises (ATX) are critical for preparing U.S. Army Combat Aviation Brigades for deployment. However, while offering the opportunity to practice mission segments at the unit level, the effectiveness of this training remains unclear due to a need for objective assessments focused on observable team behavior. Unit Commanders and trainers need tools for measuring collective task performance in order to understand performance gains, facilitate feedback, and guide the learning of aviation tactical teams. To address this challenge, a set of aviation team performance measures were developed, data were collected to validate these measures, and strategies were created to facilitate application of the measures to collective training events. The measures used behaviorally-based observations to assess performance of aviation tactical teams. The measures were used at multiple ATX events to assess performance of aviation tactical teams. Data were collected on inter-rater reliability and on agreement between the measures and overall mission performance. Results provided evidence of both acceptable reliability and validity for the measures. Moreover, requirements were developed for electronic data collection tools that can be used by unit Commanders and trainers to assess team performance at collective training exercises.

Previously, unit-level collective aviation training was accomplished through live field exercises. However, for many reasons (e.g., limited resources and lack of access to suitable practice areas), live training is less feasible than in the past. A response to these limitations was the development of the U. S. Army Aviation Warfighting Simulation Center (AWSC), a networked training system located at Fort Rucker, Alabama. The AWSC consists of 24 networked cockpit simulators that can be reconfigured to represent the Army's four currently operational combat helicopters (AH-64D Apache, CH-47D/F Chinook, OH-58D Kiowa Warrior, and UH-60 A/L/M Blackhawk). Using the AWSC, a Combat Aviation Brigade (CAB) can participate in a collective Aviation Training Exercise (ATX) that places CAB aircrews and battlestaff in a common virtual environment. While the primary purpose of ATX is to assess the readiness of battlestaff, ATX also provides an opportunity for feedback on the readiness of aircrews. The challenge addressed here is to develop methods to facilitate the provision of feedback on collective skills and task performance in a manner that meaningfully guides further development at the aviation tactical level (e.g., Company and below)

Even though individual aviation tasks are generally well defined, aviation collective tasks are comparatively poorly defined as broad mission segments that Army Aviation teams must accomplish (Cross, Dohme, & Howse, 1998). Army aviation collective tasks for reconnaissance and attack operations refer to those aviation tasks that require coordination between one aircraft and another, coordination between an aircraft (or flight of two or more aircraft) and a tactical command element (e.g., Brigade Aviation Element), and coordination between an aircraft and a Ground Commander. While tools exist to help aviators obtain step by step lists of actions to be performed, requisite underlying knowledge and skills that support aviation collective tasks cannot be inferred from such broad functions within those tasks or from task descriptions alone which lack objective performance criteria. Rather, behaviorally-anchored indicators of aviation team performance, which link observable behaviors to discrete benchmarks, should be used to evaluate performance on aviation collective tasks.

Training research (e.g., Salas, Rosen, Burke, Nicholson, & Howse, 2007; Salas, Rosen, Held, & Weissmuller, 2009; Stewart, Dohme, & Nullmeyer, 2002; Stewart, Johnson & Howse, 2007) has demonstrated that the lack of clear performance assessment criteria fails to fully exploit the effectiveness of simulation-based training events. Moreover, the military value of simulation-based training, such as ATX, is determined by performance improvement of participants within the virtual-training environment (Bell & Waag, 1998). In the case of ATX, there is a need to develop performance criteria on aviation collective tasks in order to clearly illustrate what right looks like for aircrews and leaders and to assist Observer-Controllers (OCs) in providing feedback.

The primary objective of this research effort was to develop a reliable, valid, and useful assessment system. Using this system, unit leaders and OCs could provide consistent behaviorally-based feedback to aircrews that would help distinguish high-performing teams from low-performing teams. Performance results from across training units could then be aggregated to provide unit leadership with a “snap shot” of proficiency on aviation collective tasks, resulting ultimately in better performing teams. To achieve this objective, observer-based measures of aviation performance in mission-critical collective tasks were first defined. The measures were then implemented into a hand-held electronic tablet and OCs and unit leaders rated aviation team performance in multiple ATXs. Reliability and validity analyses were conducted to identify whether the observer-based measures accurately, consistently, and appropriately predicted team performance.

Measure Development

The measures were constructed using the Competency-based Measures for Performance ASsessment Systems (COMPASS^{SM}, MacMillan, Entin, Morley, & Bennett Jr., in press) approach. COMPASS is a methodology for the development of performance measures that combines experiential knowledge of subject matter experts (SMEs) with established psychometric practices. A set of three SME-based workshops took place over the course of five months that moved from the identification of key observable behaviors to the construction of performance measures. The first and third workshops were group interviews while the second workshop consisted of individual or small group interviews. A total of 27 SMEs participated across all workshops, including 3 SMEs participated in all three workshops. SME expertise ranged from military aviators to simulation training experts and software engineers.

In the first step of measure development, the phases of the attack/reconnaissance mission were deconstructed into observable behaviors, or performance indicators (PIs), that allow an expert to recognize whether an individual or team is performing well or poorly. The resulting PIs and relevant missions/tasks provided a solid basis on which to develop benchmarked measures that are less sensitive to subjective biases and more reliable over repeated sessions. In the second step, SME-provided information was crafted into specific performance measures associated with each PI in order to create performance measures with appropriate behaviorally-based rating scales (i.e., 5-point Likert-type scales). To obtain exemplar behavior information, SMEs were asked to describe and identify explicit behaviors that were representative of good, average, and poor performance. Throughout the measure development process, care was taken to ensure that measures were operationally relevant, thorough, and appropriately worded using domain language and terminology. Altogether, 130 candidate observer-based performance measures were developed. Table 1 provides an example for the PI *Request Clearance of Fires from Ground Commander*. In the final step, SMEs were presented the full set of measures to review and revise as required to ensure the measures could be understood and accepted by a wide range of potential users. Modifications were made to the measures, resulting in a final list of 115 performance measures for assessing the performance of an aviation collective team performing an attack/reconnaissance mission.

Table 1.
Example Performance Measure - Request Clearance of Fires from Ground Commander.

Does the flight request clearance of fires from Ground Commander?				
1	2	3	4	5
Flight does not request clearance of fires		Flight considers ROE; establishes friendly/enemy positions; requests clearance of fires; not ready to effect the target while going through this process		Flight considers ROE; establishes friendly/enemy positions; requests clearance of fires; anticipates clearance and sets up shot during this process

Measures Reliability and Validity

Inter-rater reliability was first evaluated as the intended use of the measures requires that different raters use the scale similarly. After demonstrating acceptable reliability, criterion-related validity was explored to determine if measures relate to performance outcomes in aviation tactical missions. The ultimate goal of reliability and validity analyses in this effort was to evaluate how well measures performed and to inform revisions to the measures and scale anchors as appropriate.

Method

Reliability and validity data were obtained during two separate ATX events conducted at Fort Rucker, AL. A total of 21 missions across two different units were observed. Of the 21 missions, 15 were simultaneously rated by two or more experienced aviators. Three of those 15 featured three independent raters. The remaining six missions were rated by one experienced aviator. Outcome measures were obtained from 21 missions and focused on more objective outcomes of the mission (e.g., mission accomplishment, achievement of objectives, number of targets destroyed, aircraft lost). While raters evaluated flight team performance in real-time, outcomes measures were completed following the end of a mission, both collected using an electronic measurement tool. Given these data, inter-rater reliability was evaluated on the 15 missions with multiple raters in each while criterion-related validity was examined on all 21 missions.

In the absence of an existing pure criterion measure (i.e., an independent objective training or performance outcome) a substitute measure was developed. This outcome measure consisted of nine items indicating variables such as mission success, number of targets destroyed, number of friendly aircraft lost, and instances of fratricide. Given limited access to higher-level-leader raters, outcome ratings were completed by the same observers who rated the process measures. While this analysis does not speak directly to criterion validity because of rater dependencies and the absence of a true criterion, it serves as way to verify consistency and relationships between process and outcome measures.

Results

Inter-rater reliability. While inter-rater reliability is a standard approach for demonstrating that raters use measures and scale anchors similarly, evaluations of other measure properties such as percent agreement can be insightful tests of the reliability of ratings (Howell, 1997). Further, percent agreement as computed in this study can help identify measures that were especially problematic for raters to agree upon – an important step for revising as well as down-selecting the large measures set to a manageable number of the best performing and useful items. As a result, inter-rater agreement was first assessed and then followed up with a more standard inter-rater reliability analysis.

Inter-rater agreement was established using a percent agreement method based on the range of ratings for each measure across the raters (e.g., both raters within one rating point). For each level of agreement, percent agreement was calculated by dividing the observed agreement counts by the total number of possible observations. When aggregated across all rated missions, raters achieved a 72% agreement within 1-point on the Likert scales. Put differently, if one rater gave a rating of five, the other rater(s) was likely to give a rating of at least four in 72% of the occasions. Considering the many uncontrollable environmental factors present during this testing, these results are quite promising in demonstrating that raters would use scale anchors similarly.

Given the relatively high percent agreement observed in the first analysis, inter-rater reliability was computed using Cohen's Kappa (κ), a conservative measure of inter-rater agreement that accounts for chance agreement (Cohen, 1960; Fleiss, 1981). Reliability was substantial ($\kappa = 0.66$) with the 1-point-agreement threshold. Overall, the analyses suggested that different raters similarly interpreted the collective task measures. However, these results also suggested that some measures were not achieving high levels of reliability. Given these initial findings, along with the goal of refining the measures, further examination assessed which specific measures tended to have lower and higher levels of agreement.

Criterion-related validity. Only the most reliable measures were included in this analysis (i.e., rating agreement at or within 1-point in 80% of the observations). Performance measure averages were computed for each mission and were compared to average ratings for corresponding outcome measures. There was a positive relationship between performance and outcome measures such that higher ratings on performance measures were associated with higher outcome scores ($r = 0.48$, n = 32, $p < 0.05$). This result suggested that the performance measures developed to assess Army aviation collective skills do predict performance outcomes and are, therefore useful and valid predictors of performance. Taken as a whole, the results suggest that while the developed measures have some degree of validity, further work is required to refine the whole measures set prior to full implementation. Combined with reliability data, these results offer guidance on how to revise and construct the most effective measure set.

Discussion

The primary objective of this research effort was to develop reliable, valid, and useful tools to assist Leaders and trainers in assessing aviation collective performance. Using these measures, trainers are anticipated to be better able to provide consistent, behaviorally-based feedback that can help to improve the performance of aviation teams. Here, the focus was on collective tasks critical to performing typical scout/reconnaissance missions. More generally, beyond ATX, these measurement tools could also be useful in preparing for and conducting assessments in a variety of Army aviation collective training events (e.g., at home station).

The research effort reported here resulted in the construction of 115 draft measures focusing on key skills for flight teams in collective tasks. For these draft measures, initial data concerning reliability and validity were collected. These data provided evidence that the measures are in general reliable, and suggested a modest correlation between reliable performance measures and outcome measures. It should be noted that while the findings on reliability and validity were limited and preliminary, these analyses provided data on the subsets of measures that are most and least reliable, which enabled measure revision and refinement. In addition, information was collected on the requirements for tools to best enable use of the measures that will guide subsequent implementation. Based on these findings, the measures set was reduced to 105 well-performing measures, and strategies to facilitate their use were identified.

Collectively, these findings support a scientifically-based implementation plan that is designed to create a comprehensive performance assessment system. Several key objectives of this plan are to:

- Implement the refined observer-based performance measures in hand-held, tablet-based measurement tools to enable organization of measures and electronic capture of ratings for debriefing and performance tracking.
- Explore and implement related system-based measures that, once combined with observer-based metrics, could enable a more complete assessment of collective skills through leveraging of simulator data streams.
- Design and create debriefing tools that provide targeted feedback on team performance.

Ultimately, these measurement tools will enable OCs to evaluate aviation teams as they perform collective tasks at ATX. Similar evaluations by unit leaders and instructor pilots are anticipated to be possible using these tools in other collective training environments as well. Such evaluation can illuminate the status of underlying knowledge and skills and enable formative feedback that is likely to guide learning and foster development of strong teams in collective training events.

References

Bell, H. H., & Waag, W. L. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *International Journal of Aviation Psychology, 8*, 223-242.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Cross, K.D., Dohme, J.A., & Howse, W.R. (1998). *Observations about defining collective training requirements: A White Paper prepared in support of the ARMS program*. (ARI Technical Report 1075). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. ADA349437).

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley, (2nd ed).

Howell, D. C. (1997). *Statistical methods for psychology (4th ed.). Belmont, CA: Duxbury Press.*

MacMillan, J., Entin, E. B., Morley, R. M., & Bennett Jr., W. R. J. (in press). Measuring team performance and complex and dynamic military environments: The SPOTLITE method. *Military Psychology.*

Salas, E., Rosen, M.A., Burke, C.S., Nicholson, D., & Howse, W.R. (2007). Markers for enhancing team cognition in complex environments: The power of team performance diagnosis. *Aviation, Space, & Environmental Medicine, 78*, B77-85.

Salas, E., Rosen, M.A., Held, J.D., & Weilssmuller, J.J. (2009). Performance measurement in simulation-based training. *Simulation & Gaming, 40*, 328-376.

Stewart, J. E., Dohme, J. A., & Nullmeyer, R. T. (2002). U.S. Army initial entry rotary-wing transfer of training research. International Journal of *Aviation Psychology, 12*, 359-375.

Stewart, J. E., Johnson, D. M., & Howse, W. R. (2007). *Fidelity requirements for Army aviation training devices: Issues and answers*. (ARI Research Report 1887). Arlington, VA: U. S.