Developing and Evaluating Performance Measures for Manned-Unmanned Teaming

John E. Stewart
Army Research Institute
Fort Benning, GA
john.e.stewart9.civ@mail.mil

Courtney R. Dean
Aptima, Inc.
Woburn, MA
crdean@aptima.com

Troy Zeidman
Imprimis, Inc.
Huntsville, AL
troy.zeidman@imprimis-inc.com

Scott E. Graham
Army Research Institute
Fort Benning, GA
scott.e.graham.civ@mail.mil

ABSTRACT

The role of U.S. Army unmanned aircraft systems (UAS) is becoming increasingly important in tactical combat missions. Consequently, training critical skills required for manned-unmanned teaming (MUM-T) becomes more important, especially for UAS operators. In order to effectively train MUM-T skills, reliable and valid performance measures are required. Scaled observer-based performance measures can add objectivity to the process of assessing training outcomes, providing formative feedback, and tracking team progress. To this end, 36 performance measures were developed and evaluated to assess training-critical MUM-T skills. Draft performance measures were developed and refined with input from senior UAS operators and scout- attack pilots with MUM-T experience. For each performance measure, five-point behaviorally-anchored rating scales were produced representing "good," "average," and "poor" performance of the skill. The content validity of the measures and the usability of the rating scales were determined by a second group of senior UAS operators and scout-attack pilots. Most MUM-T measures were deemed relevant to the mission and observable. Six measures with low consensus by participants on relevance and/or observability were determined *not* to be practically usable. Some of these unusable measures did not reflect the role of UAS aircrews in current MUM-T operations. The measures were designed to be collected as "over the shoulder" observations. As such, a trainer, in the live or the virtual environment, could easily apply the measures. Because the resulting measures use quantitative scales that include exemplars of good-to-poor performance, they can be easily applied to unit performance assessment sessions, such as training " hot wash" and after action reviews.

ABOUT THE AUTHORS

John Stewart is Senior Research Psychologist at the U.S. Army Research Institute Fort Benning Research Unit's Fort Rucker Element. His primary area of interest is aviation psychology, which includes simulation-augmented training, performance measurement, and communication and coordination between crewmembers of manned and unmanned aircraft. He holds a PhD in Social Psychology from the University of Georgia.

Courtney Dean is a Senior Scientist and Project Manager with Aptima, Inc. specializing in training design and development, and also performance measurement. He holds an MA in Applied Psychology from the University of West Florida.

Troy Zeidman is Director of Business Development for Imprimis, Inc. He is a combat decorated, former Army Aviator who has direct experience with manned-unmanned operations. He has spent the last two years interviewing and researching manned-unmanned operations. Mr. Zeidman holds a BS in Systems Engineering from the US Military Academy, West Point.

Scott Graham is Chief of the U.S. Army Research Institute Fort Benning Research Unit. He leads a research program that is developing new and innovative training methods and assessment techniques to support the implementation Army Learning Model concepts. He holds a PhD in Cognitive Psychology from the University of Maryland.

Developing and Evaluating Performance Measures for Manned-Unmanned Teaming

John E. Stewart
Army Research Institute
Fort Benning, GA
john.e.stewart9.civ@mail.mil

Courtney R. Dean
Aptima, Inc.
Woburn, MA
crdean@aptima.com

Troy Zeidman
Imprimis, Inc.
Huntsville, AL
troy.zeidman@imprimis-inc.com

Scott E. Graham
Army Research Institute
Fort Benning, GA
scott.e.graham.civ@mail.mil

INTRODUCTION

Background

The development of objective measures of performance for aviation training is not a new concept (Stewart, 1985; Stewart, Dohme, & Nullmeyer, 1999). For the past two decades performance measures have been within the capabilities of modern flight simulators. Stewart (1994) demonstrated how system-based, automated performance measures could be predefined and captured directly from the data recording system of a high-fidelity AH-64A helicopter research simulator. These system-based measures were comprehensive (e.g., control input, pilot's head orientation and aircraft state). They were validated by senior instructor pilots, and found to correlate significantly with real-time performance ratings of a set of standard maneuver tasks. Post hoc blind rankings of output graphs of several of these maneuver tasks (e.g., roll-on landing) by the same instructors showed very high concordance.

A substantial body of research has shown that both automated and precisely scaled observer-based measures can provide objective benchmarks for assessing not only the effectiveness of simulators, but training programs as well (Benton, Corriveau, & Koonce, 1993; Nullmeyer & Rockway, 1984). Observer-based measures have been shown to be superior to the current subjective criteria used to train student pilots, such as daily flight grades and checkride grades. Empirical evidence further suggests that flight training grades alone may not be a valid predictor of future aviator performance in the operational unit. For example, Bale, Rickus, and Ambler (1973) followed U.S. Navy aviators to their fleet air groups after graduation, and confirmed that performance in flight school did not predict performance in mission-oriented skills nor the abilities required to perform successfully in the field.

Stewart, Dohme, and Nullmeyer (1999) saw a critical need for performance measures keyed to mission-relevant skills for Army aviation, in order to determine if skills trained institutionally transferred to performance in the field. At that time, the authors concluded that developing benchmarks for measuring unit level performance would be very costly and difficult. This would involve assessing such tasks as "gunnery, troop insertions, lift operations, and coordination of battle plans with other units." This statement was made prior to the advent of shared virtual environments such as the Aviation Training Exercises, which facilitated development of prototype performance measures for Army aviation collective training (Seibert, Diedrich, Stewart, Bink, & Zeidman, 2011). These measures focused on the reconnaissance, surveillance, and target acquisition (RSTA) skills trained to scout and attack helicopter aircrews.

Manned-Unmanned Teaming

Manned-unmanned teaming (MUM-T) is an aviation collective activity that requires close communication and coordination between scout-attack helicopters and unmanned aircraft systems (UAS). Typically a MUM team will consist of a Flight of two manned and one unmanned aircraft. A concrete example of a MUM-T operation is a mission in which two armed helicopters and a UAS search a roadway for improvised explosive device implanters. The UAS detects three men digging alongside the road, and a truck containing ordnance. The UAS aircrew identifies, reports, and laser-designates the target, hands it over to the one of the helicopters, which destroys it with a Hellfire missile. The second helicopter assures that no friendly forces are in the line of fire. Thus UAS aircrews must now learn to execute complex cooperative engagement skills requiring coordination with manned helicopters. UAS aircrews learn primarily intelligence, surveillance, and reconnaissance (ISR) skills during institutional (schoolhouse) training. ISR involves surveillance limited to a predesignated area, while RSTA involves active reconnaissance and target engagement. RSTA skills are trained in the operational unit. Accurate measurement of MUM-T skills requires consensus among trainers that relevant behaviors are accurately described. Cognitive and procedural skills must be mastered, team performance assessed, and feedback provided that trainees can use to

improve their performance. Trends over time are also critical to determine how long it takes MUM teams to become proficient, and how often they should practice to sustain proficiency. For this, a toolset consisting of behaviorally anchored measures of known content validity would be required. These real-time metrics must have performance descriptors relevant to the task at hand, be understood commonly by leaders and trainers, and be based upon behavior that can be observed in the appropriate setting (Sticha, Howse, Stewart, Conzelman, & Thibodeaux, 2012). The challenge to developing such measures is that MUM-T doctrine is in its infancy, so content and standards for MUM-T skills and their performance may vary from unit to unit.

Technical Objectives

The purpose and technical objectives of the current research are to: (a) identify candidate performance indicators for MUM-T, (b) from these, develop prototype performance measures that are both relevant to specific critical skills, and tied to behaviors that can be observed, (c) benchmark prototype performance measures to behaviorally anchored rating scales, (d) determine the content validity of these measures. Determining utility and validity of these metrics will rely heavily upon input from subject matter experts (SMEs), at least some of whom have had combat experience in MUM-T operations.

MEASURE DEFINITION AND DEVELOPMENT

Method

Review of previous research

The Army Research Institute (ARI) team drew upon its previous performance measurement research (Seibert, et al., 2011; Sticha, et al., 2012), which had produced prototype indicators and measures for unit-level manned and unmanned aviation training. Sticha, et al used a method similar to Air Force Mission Essential Competencies (Colegrove & Bennett, 2006), to produce a list of 20 training-critical skills for MUM-T. A total of 150 prototype performance indicators tied to those skills was also generated. All UAS training-critical skills were consistent with Army doctrine and based upon critical collective tasks that were deemed appropriate to UAS operations (e.g. identify threats; conduct cooperative engagements). A detailed description of the tie between critical RSTA tasks and derivation of current performance measures is beyond the scope of this paper. This is described in Sticha, et al. and in Stewart, Sticha, and Howse (2012). The present ARI research team identified performance indicators that represented the essential elements of the MUM-T mission and structured them in a hypothetical mission timeline. Performance indicators represent critical tasks and crew interactions occurring during a mission that require proper execution. Altogether, a total of 84 performance indicators were developed for 16 mission phases. In addition to representing specific observed behaviors and interactions between members of a MUM-T Flight, performance indicators provided a foundation for the development of performance measures.

Individual interviews

Development and definition of the present performance measures began with one-on-one interviews with SMEs to identify overt behaviors related to good/average/poor team performance. Five active duty senior UAS operators and scout helicopter pilots took part in the individual interviews. Three were senior UAS operators; two were scout helicopter pilots. All were senior enlisted or warrant officer ranks.

Interview procedure

A variety of questions were asked to obtain information describing personnel (UAS operators/pilots) most responsible for each performance indicator. The questions elicited concrete behavioral descriptors for each performance indicator. These determined the measures to develop from each performance indicator. Examples of questions asked during the individual interview: (a) what might a member of the manned-unmanned team say or do to indicate good/average/ poor performance for this indicator; (b) what would cause a UAS operator or flight team to do well or poorly on this indicator; (c) in what situations could a person be observed performing well or poorly at this performance indicator? Research team members took detailed notes throughout the interviews.

Candidate measures refinement workshop

Using information gathered during the interviews, the research team developed tentative sets of behaviorally-anchored performance measures. This was done by taking each performance indicator and the associated notes from the interviews and using behavioral anchors (i.e. verbal descriptors) that define good and poor performance. One

performance indicator could have one or more measures associated with it, and these measures could describe observable behaviors for either individual roles or the entire MUM-T Flight. After the draft measures had been developed, a focus group format workshop was held to refine the initial list of candidate measures.

Participants in the measures refinement workshop

The candidate list of measures derived from individual interviews was reviewed by 11 SMEs from a combat-experienced Air Cavalry Troop consisting of both manned and unmanned aircraft. Reviews took place during a single, four-hour workshop. SMEs were a mix of eight UAS operators (enlisted and junior NCO) and three scout-attack helicopter pilots, all warrant officers (CW3 & CW4). Participants' most recent platform operation experience included Medium UAS (e.g., RQ-7B), OH-58D and AH-64D helicopters. All had recently been deployed in combat. Most had had MUM-T experience during deployment.

Workshop procedure

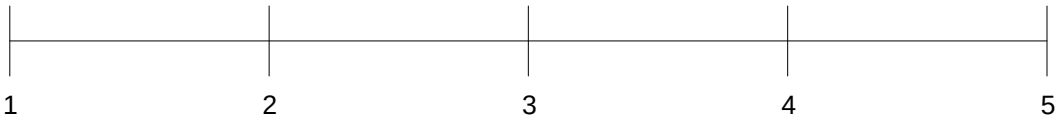
SMEs were asked to evaluate the material to ensure that performance indicators and performance measures were operationally relevant, as thorough as possible given the mission context, and appropriately worded. Each performance measure was reviewed with respect to the following criteria: (a) relevance, (b) observability, (c) measure type (e.g., scale, dichotomous, checkboxes), (d) wording, (e) scale type, and (f) scale wording. Participants were also asked if any additional measures were needed or if any should be removed completely. The result of this process was a set of measures that was developed, reviewed and refined by a mix of experts, based upon their actual combat experience in MUM-T operations.

Results and Discussion

Individual interviews

At the conclusion of the individual interviews each performance indicator was reviewed and modified to form a question that refers to the task or objective characterized by the performance indicator. For instance, using the performance indicator: "Transmit a SPOT report in accordance with SALT-W format (in accordance with TC 1-248)," the measure question was written "Does the aircrew send a SPOT report to the supported ground unit upon target detection (if required)?" Questions, scale types, and verbal scale anchors for each performance indicator were developed. The research team identified key verbal descriptors (anchors) that represented differing levels of performance. Following the identification of the three levels of performance, the identified key descriptors were formatted by the ARI research team into a draft performance measure (Figure 1). The three anchors depict varying levels of quality, or completeness. This, by design, is meant to achieve higher levels of inter-rater reliability and reduce subjectivity in ratings. The result of this effort was a total of 45 draft performance measures.

Does the aircrew send a SPOT report to the supported ground unit upon target detection (if required)?



1	2	3	4	5
Aircrew does not communicate detection of target		Aircrew announces detection to supported ground unit; does not transmit target information		Aircrew sends proper SPOT report to supported ground unit

Figure 1. Example of a behaviorally anchored rating sale (BARS) developed from interviews.

Measure refinement workshop

While many measures were modified during the workshop, some were deemed satisfactory with no edits and remained unchanged. A total of 48 measures emerged from this workshop. This list of 48 measures was culled down to 36 measures appearing in scenario form. The culling process removed those measures which were redundant or reflected behaviors that were not part of a typical mission. Several of the measures were merged where tasks and behaviors were deemed by workshop SMEs to be complementary or indistinguishable. The measure refinement workshop yielded 36 refined performance measures for MUM-T.

CONTENT VALIDATION OF MEASURES

Method

In this workshop, raters from the scout-attack and UAS communities provided input as to whether each performance measure was *observable* during a MUM-T training event and *relevant* to the MUM-T mission. These two adjectives are important criteria for the usefulness and validity of the measures. Any measure deemed "not observable" by the group would be unusable by an observer in a simulation training event. Moreover any measure deemed "not relevant" would similarly serve only as a distraction to an instructor trying to gather meaningful performance data during training.

Participants

Participants were 19 active-duty members of an Air Cavalry Troop that has deployed UAS platforms with scout aircraft in MUM-T combat missions. Nine were OH-58D pilots and 10 were RQ-7B operators. Ten were present in the morning session and nine in the afternoon session.

Procedure

For purposes of gathering utility data for each measure, a survey was administered. Each measure was rated (4 pt scale) on two criteria: relevance to MUM-T and observability (by instructor and/or trainer). Agreement was indicated by circling the preferred option (strongly agree to strongly disagree). Following the survey, a roundtable discussion was held. The follow-up discussion addressed, generally, the concept of MUM-T and inquired about experiences conducting MUM-T in combat operations. Specific inquiries referred to what challenges operators and pilots faced in integration into the Air Reconnaissance Squadron and how their roles as team members evolved from the start to the end of the tour. Of particular interest were the different experiences between the two UAS Troops as a consequence of the different integration situations (i.e. remaining with the joint UAS-manned Troop vs. being detached and placed under operational control at another level).

Results and Discussion

Overall agreement between 14 to 19 SMEs for all 36 items for both criteria was determined by computing Fleiss' Kappa, (Fleiss, 1971), a measure of categorical agreement among multiple raters. Because Kappa assumes nominal categories only, agreement was defined by the broader dichotomous categories of Disagree and Agree. For the Relevance criterion, K = .31, for Observability, K = .22. Though the distribution of Kappa is unknown, both obtained Kappas fall into the range of fair agreement (Landis & Koch, 1977). The relatively low agreements are not that surprising in light of the fact that there were two reversals in which majority of SMEs disagreed with the relevance and observability of two measures, namely, *UAS prioritizes engagement of targets* (agreement= 42%; relevance, 44% observability) and *updates engagement priority* (agreement= 42%, relevance, 39%, observability). Overall agreement across all items was 90.5% (relevance) and 87%, (observability). The Pearson product-moment correlation between aggregate ratings of the criteria of relevance and observability was significant (r = .95, df = 34, p < .001).

Table 1 presents the performance measures in descending order of consensus. While there is no simple "rule of thumb" as to what percentage of agreement is minimal for content validity and usability, it would seem reasonable to conclude that the last five measures in the list would not be usable in gauging team performance. For these, levels of relevance and/or observability are close to chance levels. The low consensus on these measures is important in that they were noted by Sticha, et al. (2012) as skills in which UAS aircrews were not currently proficient. In particular, the prioritizations of targets and deconfliction of airspace/clearance of fires are tactical skills that a UAS operator straight out of the schoolhouse would definitely not know. These measures had, at best, moderate ratings of relevance. They were, according to some SMEs, not the responsibility of the UAS aircrews. This division of opinion indicates disagreement as to whether UAS aircrews should learn to perform these skills. Overall, disagreement was low across all skills; 9.5% for relevance and 12.7% for observability, attesting to overall appropriateness and usability of most measures.

Table 1. Percent Agreement on Mission Relevance and Observability of Performance Measures

Measure	Percent Agreement	
	Relevant	Observable
UAS uses appropriate sensors	100	100
UAS recognizes threats during mission	100	100
UAS sends complete SPOT report upon target detection	100	100
UAS maintains positive identification of target after acquisition	100	100
AS reports when contact lost (last known location, direction, etc.)	100	100
Flight coordinates duties after target acquisition	100	100
UAS provides continuous reconnaissance	100	100
UAS updates target behavior (changes in size, activities, movements, etc.)	100	100
UAS provides early warnings, threat detection to supported unit	100	100
UAS follows correct procedures & format for target handover to wingman	100	100
UAS correctly identifies targets during target handover	100	100
UAS verifies location of friendly forces near target to prevent fratricide	100	100
UAS coordinates manned aircraft (launcher angles, safety fans, laser codes)	100	100
UAS uses correct procedure for remote Hellfire launch (voice/digital/laser)	100	100
Flight actively searches for target	100	94
UAS shares sensor feed with flight & communicates throughout mission	100	94
UAS updates flight on changes in Common Operating Picture	100	94
Flight incorporates ISR plan	100	89
UAS shares sensor feed with tactical operations center	100	89
UAS conducts standardized Battle Damage Assessment	95	100
UAS announces target acquisition	95	94
Flight conducts Collateral Damage Assessment	95	94
Flight confirms hostile intent prior to applying lethal force	95	94
UAS shares sensor feed with ground unit	95	71
UAS uses proper format for indirect fire mission	94	87
UAS provides updates to ground unit	89	89
Flight selects and briefs appropriate engagement scheme of maneuver	89	88
UAS proactive in executing call for indirect fire	84	81
Flight selects appropriate weapon for desired effect on target	83	81
Flight recommends (lethal/nonlethal) courses of action to ground commander	82	69
UAS proactive in airspace deconfliction in execution of indirect fires	72	67
UAS relays target direction & range to other aircraft to clear airspace	69	50
UAS acknowledges receipt (or any changes) from Fire Direct Center	68	53
UAS deconflicts airspace in preparation for missile launch	63	56
UAS prioritizes engagement of targets	42	44
UAS updates engagement priority as it changes	42	39
Note: Flight refers to the entire team, usually consisting of two helicopters and one UAS.		

GENERAL DISCUSSION

Summary and Conclusions

The resulting performance measures were behaviorally-anchored and observer-based, and most were deemed usable by SMEs. Previous work (Bink, Dean, Ayers, & Zeidman, 2014; Seibert, et al., 2011) that had developed and validated similar performance measures for Army Aviation collective training, provided useful information for the current research effort. All derived performance measures were based on critical collective tasks (Sticha, et al, 2012), as well as tactics, techniques, and procedures used in RSTA operations. As one would expect, the behaviorally-anchored measures were considerably more specific than the collective tasks and skills from which

they were derived. The final step in the development of the performance measures was focused on determining the appropriateness of their content to the specific performances expected of UAS aircrews at different levels of proficiency. We were able to identify what measures had sufficient content validity for measuring team performance. We pinpointed those skills that were measurable and developed refined measuring instruments that could be used to assess performance during MUM-T training.

Not all measures were found to be relevant to the mission and usable by leaders and trainers. Further discussion with SMEs revealed disagreement as to whether certain skills should be performed by UAS aircrews or pilots. Prioritizing and updating targets and coordinating airspace in preparation for missile launches or artillery fires showed low consensus among SMEs that these were skills appropriate to UAS aircrews. MUM-T doctrine and tactics are still evolving, so it is not surprising that the role of UAS aircrews as team members for certain skills may be unclear. It is also possible that the unit has limited time and resources to train all RSTA skills, and finds it more efficient to upgrade proficiency by building upon ISR skills which were part of Advanced Individual Training at the schoolhouse (Ingurgio & Stewart, 2014). These investigators surveyed manned and unmanned team members who had recently returned from combat. They found that the rate of communication by UAS aircrews in MUM-T combat situations most often concerned ISR aerial observation activities such as providing information about the locations of potential threats (e.g., newly patched road surface), enemy forces, and targets (e.g., mortar teams). This seems to reflect efficient role differentiation of the manned and unmanned team members. In brief, it may not be efficient to train UAS aircrews in complex RSTA skills if the crews of manned helicopters on the same three-ship team (Flight) are already performing them. The ISR skills of the UAS aircrews were expanded to incorporate new complementary roles as team members. This involved extensive socialization of UAS aircrews, whose institutional training had had a Military Intelligence focus, into Army aviation culture. A major part of this process was learning aviation-specific procedures, including call-outs and communications, as well as understanding the roles of the helicopter aircrews and how these affect their own. Much in the same way, the helicopter aircrews must become familiar with UAS capabilities and limitations, as well as the skills that UAS aircrews bring to the unit.

Finally, it is wrong to assume that all ISR skills are UAS skills and all RSTA skills are reserved for scout-attack helicopter pilots. MUM-T by definition is a team activity, and these skills are performed in concert by members who are constantly coordinating and communicating. All team members participate in the mission planning process, and for successful mission execution, each must understand the other's role, anticipate the other's actions, and communicate effectively at all phases of the mission. In short, it may be that the overarching skills which facilitate performance of the MUM-T mission are timely and effective communication between team members. UAS aircrews must understand the procedures and terminologies underlying team-level RSTA skills, even if they do not perform all of them. Analogous to a sports team, a player is responsible for knowing his or her role in the game, and those of the other players, as well as the rules pertaining to the game as a whole. To carry the analogy further, the ground rules are currently in the formative stages.

Implications

The fact that usable measures of MUM-T have been developed leads to the question of how they can be best employed. One crucial problem at present is the need for better access to training at home station (Stewart, Bruce & Dean, 2013). The skills that comprise the present set of performance measures are team-level skills depending heavily on the use of communication, coordination, and timing. If these skills are to be acquired and kept fresh at home station, this training must take place in a shared virtual environments. These performance measures were developed with the assumption that the instructors and unit trainers would use them to track performance of teams ranging from a single manned-unmanned team, to a Company-level virtual exercise. The technology for training in shared virtual environments, using networked, transportable collective training devices, currently exists and can be implemented at home station.

The major point of this paper is that the effectiveness of this training must be measured and analyzed objectively. The present research has demonstrated that the metrics for this have content validity and are usable. These observer based performance measures were designed for implementation using commercially available mobile electronics, such as tablet PCs and smart phones. Instructors would have at their disposal a means of recording and scoring observations, which could be used later at After Action Reviews and debriefings to provide useful feedback to teams as well as individual pilots and UAS operators. Performance data could be displayed graphically, providing visual aids for interpretation. By saving and reviewing aggregated data on such a mobile device, the instructor could metrically track team progress over time, and could review indications of strengths and weaknesses to plan future training sessions. This would allow emphasis on those specific skills for which the team needed additional training

or remediation. Having this kind of a database literally in an instructor's hands would also be a boon to summative and formative evaluation of training programs.

The methodology used in the present research can be described as platform-independent in that it is not confined to military aviation, but to any collective activities that teams must execute. Nor is it restricted to applications in virtual environments. Since military operations typically involve team activities, similar measures should benefit ground commanders. One example would be training ground scouts and special ground reconnaissance teams. The development of performance-measurement scales with behavioral anchors is not new, having been originated by John Flanagan in 1954. Thus the current research utilized a time-honored scaling procedure and well-established knowledge-elicitation methodology that have been shown to be effective in developing behavioral metrics. One thing that augments the effectiveness of these measures is the rapidly evolving technology that has made mobile computing available to the public. It is, in a sense, the marriage of established methodology with state of the art electronics. With these complementary technologies in hand, training experts in virtually any field requiring teamwork could prioritize training-critical skills, determine critical incidents that distinguish good and poor-performing teams, and develop objective behavioral anchors for the most important skills. This could include a variety of training environments, including industrial settings which depend upon the coordination of highly-skilled teams.

ACKNOWLEDGMENTS

Senior UAS operators and scout and attack helicopter pilots provided guidance on the measurement of skills for MUM-T. Also, dedicated personnel with combat experience from an Army Aviation unit consisting of OH-58D helicopters and RQ-7B unmanned aircraft meticulously reviewed, refined, and validated draft performance measures for content validity and utility. The result was usable metrics for assessing MUM-T performance.

REFERENCES

Bale, R. M., Rickus, G. M., & Ambler, R. K. (1973). Prediction of advanced level aviation performance criteria from early training and selection variables. *Journal of Applied Psychology, 58,* 347-350.

Benton, C. J., Corriveau, P., & Koonce, J. M. (1993*). Concept development and design of a semi-automated flight evaluation system (SAFES)*. (Technical Report AL/HR-TR-1993-1024) Brooks AFB, TX: Armstrong Laboratory Human Resources Directorate Manpower and Personnel Research Division.

Bink, M. L., Dean, C., Ayers, J., & Zeidman, T. (2014*). Validation and evaluation of Army Aviation collective performance measures*. (*Research Report No*. 1972). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Colegrove, C. M., & Bennett, W. (2006). Competency-based training: adapting to warfighter needs. (AFRL-HE-AZ-TR-2006-14). Air Force Research Laboratory: 711th Human Performance Wing.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76, 378-382*

Ingurgio, V. J., & Stewart, J. E. (2014). U.S. Army manned-unmanned teaming: examinations of critical skills. Paper presented at Association for Unmanned Vehicles International Conference, Orlando, FL May 12-15

Landis, J. R., & Koch, G. G. (1977). Measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Nullmeyer, R .T, & Rockway, M. R. (1984). Effectiveness of the C-130 weapon system trainer for tactical aircrew training. *Proceedings of the 6th Interservice/Industry Training, Simulation and Education Conference* (pp 431-440). Washington, D.C.: American Defense Preparedness Association.

Seibert, M. K. Diedrich, F. J., Stewart, J. E., Bink, M. L., & Zeidman, T. (2011). *Developing performance measures for Army aviation collective training*. (ARI Research Report No. 1943). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Stewart, J. E. (1985). Learning and performance in an air refueling part-task trainer: a preliminary data analysis. *Proceedings of the Human Factors and Ergonomics Society, 29*, 408-411

Stewart, J. E. (1994). *Using the backward transfer paradigm to validate the AH-64 simulator training research advanced testbed for aviation*. (ARI Research Report No. 1666). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Stewart, J. E., Bruce, L. L., & Dean, C. (2013). Training aviation manned-unmanned teaming skills at home station. *Proceedings of the Interservice/industry Training, Simulation and Education Conference* (I/ITSEC Paper No. 13117). Orlando, FL, December 2-5.

Stewart, J. E., Dohme, J. A., & Nullmeyer, R. T. (1999). *Optimizing simulator-aircraft mix for U.S. Army initial entry rotary wing training*. (ARI Technical Report No. 1092). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Stewart, J. E., Sticha, P. J., & Howse, W. R. (2012). What are the most critical skills for manned-unmanned teaming? *Proceedings of the Interservice/Industry Training, Simulation and Education Conference* (I/IITSEC Paper No. 12202). Orlando, FL, December 3-6.

Sticha, P. J., Howse, W. R., Stewart. J. E., Conzelman, C. E., & Thibodeaux, C. (2012). *Identifying critical manned-unmanned teaming skills for unmanned aircraft system operators*. (Research Report No.1962). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.