Integrated Performance Measurement and Assessment in Distributed Mission Operations Environments: Relating Measures to Competencies

Thomas Carolan
Micro Analysis and Design, Inc.
Boulder, Colorado
tcarolan@maad.com

Jean MacMillan, Eileen B. Entin, Rebecca M. Morley
Aptima, Inc.
Woburn, Massachusetts
macmillj@aptima.com ebe@aptima.com,

Brian T. Schreiber, Antoinette Portrey
Lockheed Martin Technology Services Group
Mesa, Arizona,
Brian.Schreiber@mesa.afmc.af.mil

Todd Denning
L3 Communications, Link Simulation and Training
Mesa, Arizona
denningte@flash.net

Winston Bennett, Jr.
Air Force Research Laboratory, Warfighter Training Research Division
Mesa, Arizona
Winston.Bennett@mesa.afmc.af.mil

ABSTRACT

Ongoing research at the Air Force Research Laboratory has underscored the importance of developing systematic methods for evaluating the impact of advanced training and rehearsal systems for both individual operators and teams. Given the expense associated with these systems, there is a critical need for reliable and valid methods and data to quantify the benefits of such systems. Previous attempts to quantify individual and crew performance have been modestly successful, but have not been integrative and inclusive. This paper describes the development of a competency-based and embedded performance measurement system for tracking individual and team performance in Distributed Mission Operations training events. The paper discusses the measurement challenges that must be addressed if focused assessment and evaluation of distributed training and rehearsal technologies and methods is to be accomplished. The paper describes the integrated approach taken by the research team at the Air Force Research Laboratory to develop and validate process- and outcome-oriented measures.

ABOUT THE AUTHORS

Thomas Carolan is a Principal Research Psychologist with Micro Analysis & Design, Inc. Boulder CO. He received his Ph.D. in Experimental Psychology from the University of Connecticut. He is involved in research and development related to training systems, performance measurement, and human performance modeling.

Jean MacMillan is Chief Scientist at Aptima, Inc. She has a Ph.D. in cognitive psychology from Harvard University. Her research has focused on simulation-based performance measurement and team training in military environments for the past 15 years.

Brian T. Schreiber is a Senior Research Scientist with Lockheed Martin at the Air Force Research Laboratory Human Effectiveness Directorate, Warfighter Training Research Division. He has an M.S. in Human Factors from the University of Illinois. He has been actively involved in military aviation research since 1994.

Todd Denning commanded the USAF Weapons School F-16 Division 1999-2001 and is now an F-16 training and employment subject matter expert for the Air Force Research Laboratory. He is an 18-year veteran weapons and tactics instructor and training director in USAF and FMS F-16 programs. Mr. Denning is an FAA-certified flight instructor and holds degrees in Electrical Engineering and Military Art and Science.

Winston Bennett, Jr. is a Senior Research Psychologist and team leader for the training systems technology and performance assessment at the Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Training Research Division, in, Mesa AZ. He received his PhD in Industrial/Organizational Psychology from Texas A&M University in 1995.

Integrated Performance Measurement and Assessment in Distributed Mission Operations Environments: Relating Measures to Competencies

Thomas Carolan
Micro Analysis and Design, Inc.
Boulder, Colorado
tcarolan@maad.com

Jean MacMillan, Eileen B. Entin, Rebecca M. Morley
Aptima, Inc.
Woburn, Massachusetts
macmillj@aptima.com ebe@aptima.com,

Brian T. Schreiber, Antoinette Portrey
Lockheed Martin Technology Services Group
Mesa, Arizona,
Brian.Schreiber@mesa.afmc.af.mil

Todd Denning
L3 Communications, Link Simulation and Training
Mesa, Arizona
denningte@flash.net

Winston Bennett, Jr.
Air Force Research Laboratory, Warfighter Training Research Division
Mesa, Arizona
Winston.Bennett@mesa.afmc.af.mil

INTRODUCTION

Distributed Mission Operations (DMO) training provides fighter pilots with opportunities to exercise tactical and teamwork skills in an advanced distributed learning environment. These opportunities to train as a team are critical for exercising the higher-order competencies required for complex combat operations.

Ongoing research at the Air Force Research Laboratory (AFRL) has emphasized the importance of developing systematic methods for evaluating the impact of advanced training and rehearsal systems for individual operators and teams. A significant requirement for continuous improvement and maintenance of proficiency is an evaluation process that can identify proficiency levels on core competencies. Important issues include defining the quantity and quality of practice necessary to achieve effective training performance and adapting training opportunities to focus on the identified needs of individual pilots and teams, challenge appropriate competencies, and maximize learning.

Research on team performance (e.g., Cannon-Bowers, Tannenbaum, Salas, & Volpe, 1995) has emphasized the role of task and teamwork competencies in effective team performance. As Figure 1 illustrates, if we understand the skill and knowledge competencies required for effective performance on mission tasks and how training characteristics impact those competencies, then closing the loop between training and readiness requires methods for evaluating competencies based on measured performance on each training event.

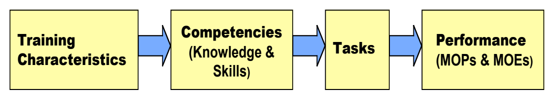

Figure 1. Role of competencies in performance

As discussed in previous papers, researchers at AFRL have developed methods for creating competency-based scenarios to support effective training and rehearsal (Bennett & Crane, 2002). This competency-based approach ties scenario events to the Mission Essential Competencies (MECs) developed for aircrew training performance (Colegrove & Alliger, 2002). The MECs identify the critical knowledge and skills necessary for successful air combat, and provide a framework for defining and measuring knowledge and skill competencies. Defining scenario events that are predictable and structured around MEC-related training objectives provides a basis for repeatable event-driven measurement of team or individual performance (Dwyer, Fowlkes, Oser, Salas, and Lane, 1997). As Bennett & Crane (2002) put it, "the instructor knows for any given moment in a scenario what competencies are being tapped, what objectives are being trained, what trigger events are about to occur, and what behaviors are critical to mission success."

AFRL continues to develop and test tools to evaluate changes in Aircrew knowledge and skill competencies that are exercised during DMO training sessions. The goal is to develop a measurement approach that integrates performance information from multiple sources including data captured from the F-16 four-ship simulation environment and data from rating scales completed by expert observers during training sessions. An integrated measurement approach will serve multiple purposes. These include providing a rich source of performance information to the pilots to support post-exercise diagnosis, developing historical performance data to evaluate training systems and strategies, and identifying strengths and weaknesses in the knowledge and skills necessary for successful air combat so that training can be focused on addressing identified deficiencies. After briefly describing the Air Superiority Mission Essential Competencies, the remainder of this paper discusses some of the approaches being taken to develop observation and simulation-based measures that are linked to the Mission Essential Competencies.

Competencies Required for Successful Performance

Mission Essential Competencies have been described as "higher-order individual, team, and inter-team competencies that a fully prepared pilot, crew or flight requires for successful mission completion" (Colegrove and Alliger, 2002). Examples of the MECs for Air Superiority include, "Detects factor groups in area of responsibility" and "Intercepts and targets factor groups."

The MEC structure developed for Air Superiority defines competencies at two additional levels of analysis necessary for specifying training objectives and design features for scenarios and the performance requirements for meeting those objectives (Bennett & Crane, 2002). Supporting competencies are high-level competency areas, such as situational awareness, communication, and decision-making that are required for proficiency in the operationally grounded MECs. Underlying proficiency in the MECs and supporting competencies are more specific knowledge and skill requirements. Thirty-two knowledge and skill competencies were identified that support effective performance in Air Superiority missions. Example knowledge requirements include "Understands threats, their capabilities, and their tactics;" "Knows criteria for commit decision," and "Understands formation standards." Examples of skill requirements include "Builds picture," "Controls intercept geometry" and "Selects tactic".

The Air Superiority MEC structure provides a methodology for defining and assessing mission readiness by relating the high level AF Task List to more detailed Training Task Lists and associated performance requirements through the knowledge and skill categories (Colegrove and Alliger, 2002). The MEC structure provides a framework for developing performance measures that are sensitive to proficiency changes in the knowledge and skill components of performance and for identifying and targeting gaps in the essential competencies required for mission readiness.

Measurement to Support Adaptive Competency-based Training

Many of the knowledge and skill elements have an impact on performance elements across all phases of a mission. Most performance requirements involve a range of knowledge and skill elements combining to produce effective results in each MEC area. Evaluating competencies in this highly complex and dynamic domain will require coordinated measurement drawing on performance information from multiple sources including the simulation environment, expert observers and the pilots themselves, and focused at multiple levels of analysis.

One performance measurement distinction that is particularly relevant to training and assessment is that between Measures of Effectiveness (MOEs) and Measures of Performance (MOPs) (e.g., Smith-Jentsch, Johnston & Paine, 1998). As effectiveness measures, MOEs typically focus on the *outcomes* of performance. As diagnostic measures, MOPs focus on performance *process*. Measures that address performance process can provide diagnostic information about why a particular outcome occurred. By assessing performance process in addition to performance outcomes, training interventions, such as post-exercise debrief and the design of training events, can be focused on identified process weaknesses.

Within the air combat DMO training environment, MOEs and MOPs provide quantitative measurements that provide a basis to describe proficiency levels for MECs. For example, MOEs for air combat involve the variables used to measure mission effectiveness as a whole. These are measures such as kill ratios, bombs on target, fratricides, and mortalities. At another level of analysis MOEs can also be used to describe outcomes of particular MEC phases. For example, number of bandits detected might be described as a measure of effectiveness for "detecting factor groups in area of responsibility." However it provides no diagnostic information to guide training.

By defining MOPs in terms of the knowledge and skill elements that contribute to task performance, MOPs can serve as a link between training, essential knowledge and skills and performance outcomes. Individual and teamwork competencies represent the knowledge and skill areas that generate performance and that advanced training is designed to exercise and strengthen in the context of applied situations. In the air combat environment MOPs measure processes such as how successful individual pilots are at weapons engagement zone (WEZ) management, radar use, crank angles, notch mechanics and communication. Training in DMO environments should improve performance process, while also improving mission outcomes such as increased shot and kill ratios and decreased fratricides and mortalities. Focusing training on improving performance process, and using MOPs to assess training needs should improve performance outcomes over time.

While outcome measures can usually be objectively quantified, process measures often involve some combination of objective information and subjective assessments. Where the task or skill consists of observable behaviors that can be compared to well defined performance requirements, quantitative measures are developed. The most basic levels of measurement involve capturing information to identify what happened and developing quantitative measures related to timing and accuracy by comparing relevant parameters to specified performance requirements.

DMO training exercises are essentially simulation-based training and the simulation environment provides a rich source of data for developing precisely defined quantitative measures of performance. Many air combat skills can be measured and assessed using simulator-based performance data. Performance assessments that are the best candidates for automation involve skills that have well defined performance standards or known optimums. Examples of the first include CAP contracts and formation standards that specify acceptable distance, speed, and position parameters. An example of the second is the notch, a defensive skill that has a defined optimum that serves as a basis of comparison to a pilot's notch performance under appropriate conditions. Predominately cognitive skills such as assessing a tactical situation present a challenge for automated evaluation techniques and may require inferential or probabilistic approaches. While cognitive skills are typically assessed by expert raters using observable outcomes and well designed measurement tools, simulation-based data can also be applied to support expert assessments.

Observed performance in any given situation may rely on a number of different skill and knowledge elements. The ability to assess performance in terms of the underlying competencies may require combining observation-based and simulation-based measures. Assessing the strength of knowledge based supporting competencies such as situational awareness has typically involved techniques that query the training participant directly.

Developing performance measures that provide indicators of individual and team proficiency levels on knowledge and skill competencies presents a challenge for both simulation-based and observation-based measurement techniques. For simulation-based approaches one challenge is to provide a structure to constrain and organize the many possible performance variables that can be captured, measured, and used to inform assessments. Finding the right balance between a focus on selective measurement to support current training objectives and exploiting the capability to continuously monitor and assess data relevant to multiple performance requirements is a factor in the effective use of automated performance monitoring and assessment capabilities. The MEC structure provides an organizing framework to guide the capability to monitor and assess many performance requirements in each exercise. Mapping performance requirements to knowledge and skill competencies creates the potential to continually update the evaluation of individual and team competencies, and to use the resulting profiles to refine training decisions.

A major challenge for developing observation-based performance ratings is deciding where to focus the observer's attention for measurement. Assessing any one of the knowledge and skill areas could occupy the resources of an observer during a training session. The MEC structure provides a basis for structuring scenarios so that observation-based measurement can be focused on specific, observable behaviors that are linked to specific events and competencies. In order to achieve the goal of evaluating MECs, we are developing multiple measurement approaches, each with a different and complementary primary focus but with enough overlap to assure convergence of measures, methods, and tools into an integrated capability to monitor performance and assess competencies.

Two of these approaches are highlighted in the following sections. SPOTLITE is a method for focusing the development of performance measures and is applied to an observation-based measurement instrument developed for the DMO training environment. PACES emphasizes the development of a performance requirements-based assessment capability and is applied to the development of an automated performance and competencies assessment tool. The Performance Effectiveness and Evaluation Tracking System (PETS) provides the core simulation-based performance data tracking and measurement system for the AFRL DMO environment and is briefly discussed.

DEVELOPMENT OF OBSERVER-BASED PERFORMANCE ASSESSMENT METHODS

Measures for assessing performance in the complex DMO simulation environment must provide insight into whether learners are acquiring the competencies needed for success in the environment, and whether they possess the knowledge and skills that underlie these competencies, as discussed above. Given a definition of the competencies, knowledge, and skills required for successful performance, the challenge becomes one of selectively focusing the measurement process. Typically, a complex multi-person simulation-based training environment offers a myriad of behaviors that *could* be measured, and far more data than can be easily processed or understood. Further, there is rarely one "right answer" or only one path to success in a simulated situation, and multiple team members contribute in varying degrees to the outcomes that are observed.

The SPOTLITE method (MacMillan et al., 2002), developed under the sponsorship of AFRL, addresses the difficult problem of focusing performance measurement in a complex multi-person simulation-based training environment. The goal is to develop feasible, reliable measures that are sensitive to variability in performance, diagnostic of performance difficulties, and validated by their relationship to overall outcomes. The method was designed and first applied to develop a measurement instrument for four-person teams of F-16 pilots training for air-to-air combat in a high fidelity DMO simulation environment. Figure 2 illustrates the steps in the SPOTLITE measure-development process.

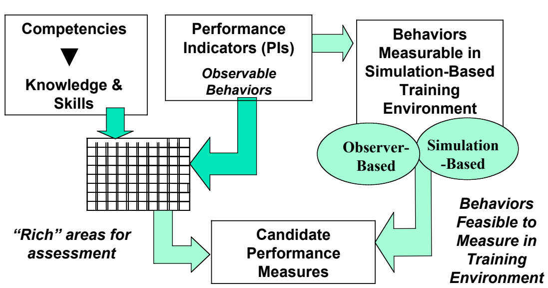

Figure 2. Steps in Measure Development with the SPOTLITE Method

The first step, once competencies, knowledge, and skills have been defined, is a knowledge-elicitation session with subject matter experts (in this case, F-16 Instructor Pilots) in order to identify a series of behavioral *Performance Indicators*. These Performance Indicators describe *observable behaviors* that are associated with the Mission Essential Competencies. Performance Indicators are focused around specific types of missions, and were elicited by asking the Instructor Pilots to describe behaviors that they had frequently seen as instructors that indicated to them that a pilot (or a team of pilots) were at a low level of readiness and needed more experience and instruction, or were at a high level of readiness. In the sessions we conducted, we found that the Instructor Pilots were easily able to generate concrete, specific descriptions of behaviors that indicated low levels of readiness (e.g., does not have radar properly set up, uses extraneous communication). Interestingly, the instructors most easily defined high levels of readiness as being the absence of commonly occurring errors, rather than as the presence of defined expert behaviors. This may be specific to the fighter pilot environment, however, which is an extremely demanding one in which a very high degree of expertise is needed to simultaneously perform multiple tasks at acceptable levels.

Given a list of skills and knowledge (based on Mission Essential Competencies) and a list of Performance Indicators (specific behaviors indicating readiness) we generated a blank matrix and asked subject matter experts to fill in the cells of the matrix by indicating the importance of each of the types of skills and knowledge in supporting the behaviors described by the Performance Indicators. Analysis of these data provided us with a method for identifying the *"rich" areas for assessment*—the observable behaviors that draw most extensively on the knowledge and skills underlying the Mission Essential Competencies. In a simulation-based training environment, there are typically many things that can be measured. This approach allowed us to identify those behaviors and points in the scenario where measurement will yield the most valuable diagnostic results. The process provided us with an understanding of what *should* be measured for F-16 air combat missions.

Not all behaviors that can be observed in the real world can be seen in a simulation-based training environment. In parallel to identifying the promising areas for measurement, we conducted an examination of the simulation environment to determine what *could* be measured in that environment. This analysis was conducted through multiple sessions with subject matter experts using playback of training sessions to identify behaviors actually seen with F-16 four-ship teams training in the DMO facility. We selected several "benchmark" scenarios that were used in almost all training sessions and reviewed the performance of multiple teams in each of these scenarios to identify frequently observed behaviors in the specific DMO environment of interest.

Two types of measure are potentially available: (1) data that can be taken directly from the simulation (e.g., taking shots that were within weapons range); and (2) data that can be obtained by observation during the training sessions or from observation of playback of session recordings (e.g., whether the flight lead made a timely "sorting" call to other team members). The combination of what *should* be measured based on the desired competencies and what *could* be measured in the training environment provided us with the basis for generating an initial set of performance measures. The goal is to obtain data directly from the simulation wherever possible, reserving the expert judgment and necessarily limited attention of observers for assessing those behaviors that are not easily assessed through automated means. Based on this analysis, we developed an assessment rating form with behaviorally anchored ratings scales to be used by observers during a training session.

The final step in measure development is applying, assessing and improving the measures. Figure 3 illustrates this process. In assessing measures, we consider:

1. *Sensitivity.* Does the measure distinguish between multiple performance levels for the target population, or does everyone score at the bottom of the scale (floor effect) or at the top of the scale (ceiling effect)?
2. *Reliability.* For measures that are based on observation, do multiple observers rate the same behavior in the same way? For simulation-based measures, reliability is not an issue.
3. *Validity.* Do the detailed behaviors being measured relate to overall performance in the expected direction? A useful technique for validity testing is to have several extremely knowledgeable observers independently rate the overall performance of an individual or team. If these observers agree, then the correlation between this overall rating and the more detailed measures can be used to validate the detailed measures. This "convergence of experts" validation technique was pioneered in a study of the decision-making expertise of Army battle commanders (Serfaty, MacMillan, Entin, and Entin, 1997).

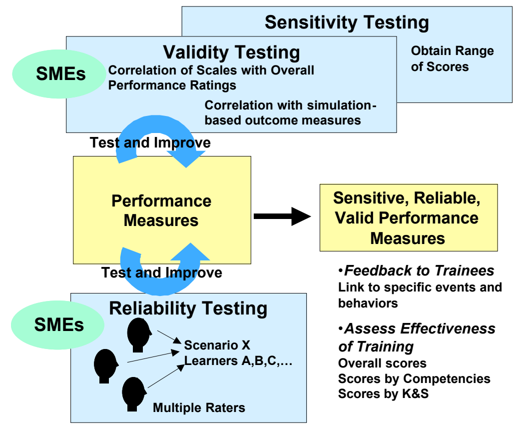

Figure 3. Assessment of Measure Sensitivity, Reliability, and Validity

The final outcome of the SPOTLITE measure-development process will be an observer rating form, to be in used in conjunction with data available directly from the simulation, that will produce a valid, reliable, sensitive measures of the performance of F-16 four-ship teams during DMO training sessions.

DEVELOPMENT OF SIMULATION-BASED PERFORMANCE ASSESSMENT METHODS

The air-to-air MECs encompass a vast array of skills and competencies, each of which may require capturing multiple variables. The sum of these leads to thousands of individual variables to monitor related to relevant initial state situations, process behavior skills, and mission outcome measures. This requires identifying and capturing all variables relating to MECs, and involves monitoring, time-stamping, and storing thousands of variables from different sources. Once the variables are identified and defined, the difficulties then lie in measuring the variables with data from a complex networked simulation that does not lend itself readily to performance monitoring.

Developing an Automated Performance Effectiveness and Evaluation Tracking System

The initial need to create an automated objective measurement tool to assess both higher- and lower-level F-16 air combat MECs in the DMO training environment led to research and development at the Air Force Research Laboratory in Mesa, AZ. One result of this research is a proof-of-concept automated distributed performance effectiveness and evaluation tracking system (PETS). This performance evaluation tracking system collects data from the network and uses it as inputs to the different algorithms needed to generate the quantitative measurements for analyses.

The development of PETS involved workshops and usability trials with subject matter experts, researchers and software engineers to identify the many variables necessary in order to measure specific MEC-driven metrics and their associated MOEs and MOPs. The PETS system underwent a comprehensive testing, refinement, and, validation process to ensure that the hundreds of unique measures were being collected accurately (see Schreiber, Waltz, Bennett & Portrey, 2003 for details).

PETS provides a unique capability for subject matter experts and researchers to identify, define and validate new MOEs and MOPs to quantify elements of performance that are inadequately captured, or not captured at all, by current measurement systems. PETS development continues to increase its value as both a performance feedback tool for pilots and as a training evaluation and research system. In addition it provides a rich source for performance information that can be used in the development and evaluation of performance assessment and diagnostic tools to support adaptive training.

Developing Performance and Competencies Requirements-based Evaluation

As discussed previously, the goal of performance measurement is to determine if pilots are acquiring the knowledge and skill competencies needed for success in the combat environment. Simulation-based performance data provides objective evidence that can be applied as input to individual proficiency estimates for the knowledge and skill competencies. The objective of the Performance and Competencies Evaluation Support (PACES) tool is to implement a framework for assessing the changes that develop over DMO training sessions in the Aircrew knowledge and skills necessary for successful air combat. The approach is designed to use both automated and observer generated performance data as evidence for the strength or weakness of particular competencies.

To simplify and structure the automated performance evaluation problem, the assessment approach has focused on comparing observed performance to an accepted solution or performance requirement. This first approximation strategy for developing competency profiles starts by identifying conditional relations between specific performance requirements and the relatively high-level knowledge and skill elements. At this level the goal is to identify which knowledge and skill elements are required for effective performance and the relative impact of each on success or failure of the task.

Performance Requirements Analysis

The Mission Essential Competencies are defined based on how warfighters execute the Find-Fix-Track-Target-Engage-Assess sequence in the combat operating environment (Colegrove and Alliger, 2002). A team of expert pilot trainers have been involved in an intensive workshop approach to defining the task and performance requirements that constitute each Mission Essential Competency. The analysis process starts with a decomposition and specification of individual and team tasks involved in the mission phase associated with each MEC. The performance requirements analysis process populates a spreadsheet with the MEC related tasks defining the rows and the measurement properties to be identified for each MEC subtask defining the columns of the spreadsheet grid. This decomposition is then used to populate the Performance and Competency Evaluation Support (PACES) tool's database, providing the framework needed to support performance evaluation. Figure 4 illustrates part of the task requirements identified for MEC 1.

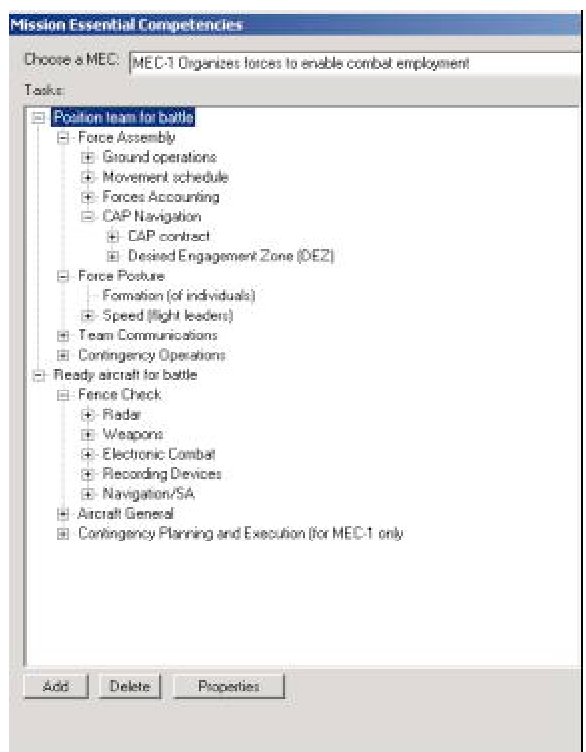

Figure 4. MEC 1 Decomposition Examples

Many of the MEC 1 tasks can be measured objectively, in some cases using performance data captured from the simulation environment and in other cases by observers listening to network communications. This decomposition process produced over 300 cognitive and behavioral performance items for MECS 1 through 4.

The next step was to identify performance measures for each item and specify how to evaluate each measure. Many of the tasks and subtasks have objective, observable components or consequences that can be captured using data from the simulation environment or by observers. The hierarchical structure provides a framework that allows an evaluation of higher level tasks based on whatever information is available from the lower level measures. Our goal therefore is to build a comprehensive performance requirements database and implement those measures that are feasible given the available performance data.

We identified the probable source of the evaluation data for each measure as either simulation-based or observer-based and made an assessment as to the current or likely availability of performance data to support each measure. For each measure, performance standards were identified, as available, and the source of regulatory guidance for each standard was specified. The standards and the briefed plan provide the basis for defining expected performance. To support performance evaluation, information about when to measure each performance item is needed. Start and stop triggers were specified for each measure. The final item in the performance requirements portion of the spreadsheet defines who should be evaluated for each task. This can be all fighters, wingmen only, element leads only, flight lead only, or any individual fighter. In addition we are identifying measures that are defined at the team, rather than individual, level. Figure 5, illustrates properties associated with one measure, CAP radial contract.

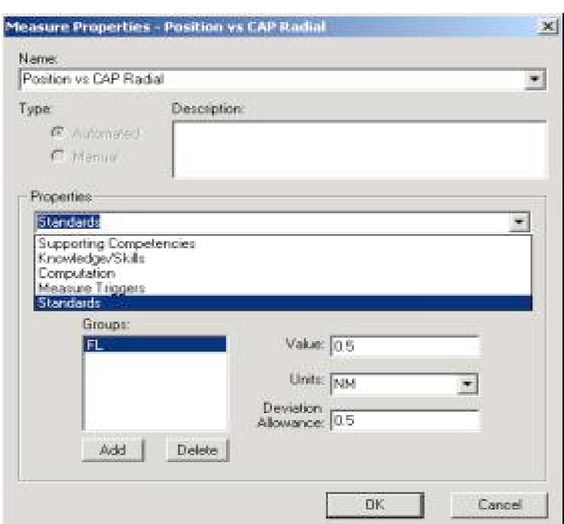

Figure 5. Properties associated with each measure

This approach allows us to (1) define outcome and process measures for each MEC, (2) identify knowledge and skill competency requirements for effective performance on each measure, and (3) identify those measures that can be used to support evaluation of team level competencies.

Competency Requirements Analysis

After the performance requirements are defined the next step is to identify the knowledge and skill competencies required for successfully achieving the performance criteria defined for each measure. The method we used to identify the competencies required for each measure was to have the team of expert pilot instructors work through the task requirements defined for each MEC and discuss the process for accomplishing each task, the information and knowledge sources they would use, and the component skills that go into each task. This cognitive replay and analysis process led to agreement among the experts on the knowledge and skills that were involved for each performance requirement. The performance and competencies requirements spreadsheet (Figure 6) was used to guide the analysis and capture the results.

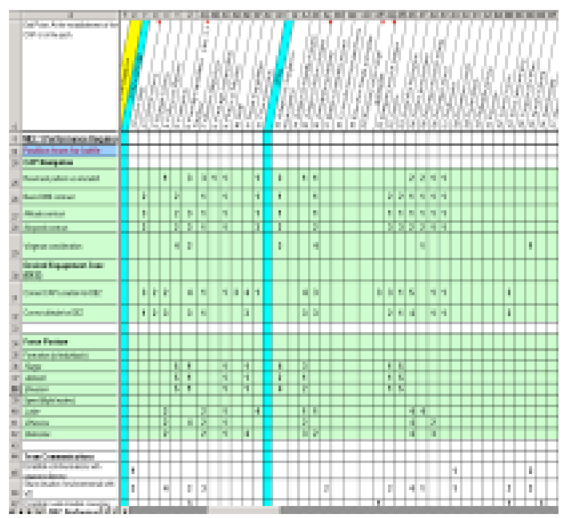

Figure 6. Identifying Competency Requirements

As Figure 6 illustrates, the expert pilot instructor team assigned a weight to each measure-by-competency cell representing the importance of that knowledge/skill to effective performance. The values on the importance weight scale ranged from 1 to 5, with 5 being the most important for success. The weight values are anchored by a descriptive definition that indicates the importance of the knowledge or skill to successful performance. The anchors used for each scale value are given in Table 1.

Table 1. Weight values and associated descriptions.

Weight	Description
5	**Knowledge or skill essential for <u>initiation, guidance, and completion</u> of a task** -- Provides set of K or S elements required to select and prioritize task within mission essential competency context. -- Lack of this skill leaves the operator with no ability to recognize, understand, or execute what is required for the given situation.
4	**Knowledge or skill essential for <u>basic task accomplishment</u>.** -- Provides set of K or S elements required to carry out execution steps for desired outcome. -- Lack of this area leaves the operator without clear direction on required actions.
3	**Knowledge or skill contributes to <u>successful outcome</u>.** -- Provides discriminator K or S that drives task to successful end state. -- Lack of this area establishes equal probability of success / failure.
2	**Knowledge or skill enhances <u>effectiveness</u>.** -- Contributes to degree of task success and impact on tactical outcome. --Lack of this area may degrade outcome; an issue of notable performance enhancement.
1	**Knowledge or skill enhances <u>efficiency</u>.** -- Makes task go smoother, better coordinated, less time/energy expended. -- Lack of area is highly unlikely to decide success or failure; an issue of fine tuning.

The knowledge and skill weights are assigned to each performance requirement in the context of working through the cognitive and behavioral steps involved in carrying out each phase of a mission and responding to ongoing events. Once this step is completed for each MEC, the expert pilot instructor team then reviews each knowledge and skill competency, looking at the pattern of weights assigned to the competency, and adjusts the weights by comparing the relative importance of the competency over the set of measures. The result is a mapping from measures to competencies that captures the important performance indicators for each competency while also providing a basis for relating more detailed performance-requirements-based measures to underlying competencies.

The weighted mapping from measures to competencies provides a means for estimating the strength of particular competencies based on the pattern of performance within and across MECs (Carolan, Shurig & Bennett, 2003). Some measures draw primarily on a particular knowledge or skill; others require the contribution of multiple competencies for success. In mapping measures to required competencies we found that a finer grained analysis of knowledge and skills would improve the precision of the mapping in many cases. For those competencies that are defined more broadly, different measures may be tapping substantially different aspects of the knowledge area or skill. The expert pilot instructor team decomposed the knowledge and skill categories down to a finer level of analysis and reassigned the competencies and weights to each measure at this lower level. This additional level of detail supports a more diagnostic level of assessment aimed at relating competencies to outcome and process measures.

One strategy we are investigating to support automated performance measurement for complex behaviors is the use of expected performance models that specify appropriate task performance under various conditions based on well defined tactical decision making logic. Where absolute performance standards are not appropriate, this strategy will allow an assessment algorithm to assign ratings based on performance relative to the dynamic standard defined by model performance. In addition, assessment of performance on each of the critical variables will provide a more diagnostic evaluation.

Plans for sensitivity and validity evaluation and refinement of the measures involve comparing the output of automated measures (1) over a set of test scenarios flown to emulate pilots with different skill levels, and (2) against the ratings of expert instructor pilots evaluators for a number of engagements. Test scenarios have been flown specifically for evaluation and development of the automated measures.

INTEGRATING MULTIPLE MEASURES

A primary purpose of performance measurement is to identify strengths and weaknesses in the knowledge and skills necessary for successful air combat so that training can be focused on addressing identified MEC deficiencies. Assessing knowledge and skill proficiencies based on performance data can be thought of as assigning "credit or blame" to a knowledge or skill element or combination of elements for observed performance deficiencies. The goal is to develop individual and team competency profiles based on performance over a single DMO exercise and a series of DMO exercises. The competency profiles can then be used to track progress and tailor exercises based on individual and team mastery or lack of mastery of specific competency areas (Bennett, Schreiber & Andrews, 2002).

This evaluation process combines objective performance information automatically generated using training simulation data files with performance information generated by instructor/observers. Another potential source of information to support competency assessments is data provided by the pilots themselves, either during debrief or through post-exercise questionnaires. This source of information may be particularly useful in assessing knowledge competencies. A range of knowledge and skills combine to produce effective performance in any MEC area. Objective simulation-based measures and observation-based measures together provide a rich basis for assessment of the knowledge and skills that support each MEC. By using a common measurement framework, observation- and simulation-based data can be integrated to provide assessments at the knowledge, skill, and MEC level.

REFERENCES

Bennett, W., Jr., & Crane, P. (2002). The deliberate application of principles of learning and training strategies within DMT. Paper presented at the NATO SAS-038 Working Group Meeting, Brussels, Belgium.

Bennett, W., Jr., Schreiber, B.T., & Andrews, D.H. (2002). Developing competency-based methods for near-real-time air combat problem solving assessment. Computers in Human Behavior, 18 (773-782).

Cannon-Bowers, J. A., Tannenbaum, S. I., Salas, E. & Volpe, C. E. (1995). Defining team competencies and establishing team training requirements. In R. Guzzo & E. Salas (Eds.), *Team effectiveness and decision making in organizations* (pp. 333-380). San Francisco: Jossey Bass.

Carolan, T. F., Shurig, I., Bennett, W. Jr., (2003). Evaluating Competencies: A Probabilistic Approach. Paper presented at the Behavioral Representation in Modeling and Simulation conference, Scottsdale, Arizona.

Colegrove, C.M., & Alliger G.M. (2002). Mission essential competencies: Defining combat mission requirements in a novel way. Paper presented at the NATO SAS-038 Working Group Meeting, Brussels, Belgium.

Dwyer, D.J., Fowlkes, J.E., Oser, R.L., Salas, E., & Lane, N.E. (1997). Team performance measurement in distributed environments: The TARGETS methodology. In Brannick, M.T., Salas, E., & Prince, C. *Team Performance Assessment and Measurement: Theory, Methods, and Applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

MacMillan, J., Littleton, E.B., Price, J., Miescher, S.C., Entin, E.E., Gentner, F.C., Tiller, T.C., & Cunningham, P. (2000). A Streamlined Scenario-Based Methodology For Evaluating Distributed Team Performance, SBIR Phase I Final Report. Woburn, MA: Aptima, Inc.

Serfaty, D., MacMillan, J., Entin, E.B., & Entin, E.E. (1997). The decision making expertise of battle commanders. In G.A. Klein & C.E. Zsambok (Eds.) *Naturalistic Decision Making*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schreiber, B.T., MacMillan, J., Carolan, T. & Sidor, G. (2002). Evaluating the Effectiveness of Distributed Mission Training Using "Traditional" and Innovative Metrics of Success. Paper presented at the NATO SAS-038 Working Group Meeting, Brussels, Belgium.

Schreiber, B.T., Waltz, E., Bennett, W., Jr., & Portrey, A. M. (2003) Development of a Distributed Mission Training Automated Performance Tracking System. Paper presented at the Behavioral Representation in Modeling and Simulation conference, Scottsdale, Arizona.