

## DEVELOPING OBSERVER-BASED MEASURES FOR ASSESSING THE EFFECTS OF ADVANCED TECHNOLOGIES ON CREW RESOURCE MANAGEMENT

Amy L. Alexander, Jamie L. Estock, Jeff Beaubien, & Jon Holbrook  
Aptima, Inc.  
Woburn, MA

Previous research has shown that up to 80 percent of all commercial aviation accidents are the result of human error, including inadequate decision making, ineffective communication, inadequate leadership, and poor task or resource management. Currently, a number of “smart” flight deck technologies – such as Synthetic Vision Systems (SVS) and Enhanced Vision Systems (EVS) – are being developed to prevent, intervene, and/or mitigate pilot error. In some instances, these technologies are essentially acting as an additional crewmember, thus changing the dynamics of crew interaction on the flight deck. The specific effects of these advanced technologies – both positive and negative – on crew resource management (CRM) performance are difficult to quantify. Performance measures that are sensitive to technology insertion must be developed to determine these impacts. To address this issue, we are developing observer-based measures for assessing the effects of new technologies on CRM performance. This paper focuses on the systematic process used to develop these performance measures.

### Introduction

Although commercial aviation is often cited as the safest mode of transportation, the relative fatal accident rate has remained fixed over the past three decades due to an overall increase in air travel and accidents per year (Flight Safety Foundation, 2005). Previous research has shown that up to 80 percent of all commercial aviation accidents are the result of human error (Boeing, 2005). The underlying causes of these errors are many: inadequate decision making, ineffective communication, inadequate leadership, and poor task or resource management (Cooper, White, & Lauber, 1980; Helmreich, Merritt, & Wilhelm, 1999). Currently, a number of “smart” flight deck technologies – such as Synthetic Vision Systems (SVS) and Enhanced Vision Systems (EVS) – are being developed to prevent, intervene, and/or mitigate pilot error in the cockpit. In some instances, these technologies are essentially acting as an additional crewmember, thus changing the dynamics of crew interaction on the flight deck.

The insertion of new technology on the flight deck will necessarily impact flight-related operations and crew functioning. The specific effects of these advanced technologies – both positive and negative – on crew resource management (CRM) performance are difficult to quantify. Although a number of observer-based CRM measures exist (e.g., the University of Texas Line/LOS checklist, the Approach and Landing Accident Coding Form, the European NOTECHS system, the CRM Assessment System Expert Tool), they focus on general CRM issues, not on the application of CRM with regard to emerging technologies. Performance measures that are sensitive to technology insertion must be developed to determine these impacts. To address

this issue, we developed observer-based measures for assessing the effects of new technologies on CRM performance. Although the current paper only discusses the development of observer-based measures, both self-report and system-based measure development is considered an essential next step in providing a comprehensive view of crew performance.

This paper focuses on the process used to develop observer-based, technology-sensitive CRM-related performance measures. Based on a review of the literature and recent accident statistics, we focused the development of performance measures within the context of SVS and EVS technologies. These technologies are being designed to reduce the occurrence of low-visibility induced accidents, including controlled flight into terrain (Alexander, Wickens, & Hardy, 2005; Prinzel, Comstock, Glaab, Kramer, Arthur, & Barry, 2004; Schnell, Kwon, Merchant, & Etherington, 2004). These systems provide a real-time representation of the outside world along with advanced symbology to support guidance and control. NASA has been conducting research on the design, development, and implementation of SVS/EVS technologies for several years.

### Method/Results

The development of performance measures sensitive to the insertion of advanced technologies involved a systematic process consisting of five steps, as shown in Figure 1. These steps included: 1) defining the CRM skills pilots need to interact effectively with advanced technologies, 2) identifying performance indicators, or observable behaviors, that allow an expert rater to recognize whether the crew is

performing well or poorly on CRM skills, 3) identifying behaviors measurable in a simulation-based environment, 4) developing an initial set of

candidate performance measures, and 5) assessing measure sensitivity, reliability, and validity. Each of these steps is described below in more detail.

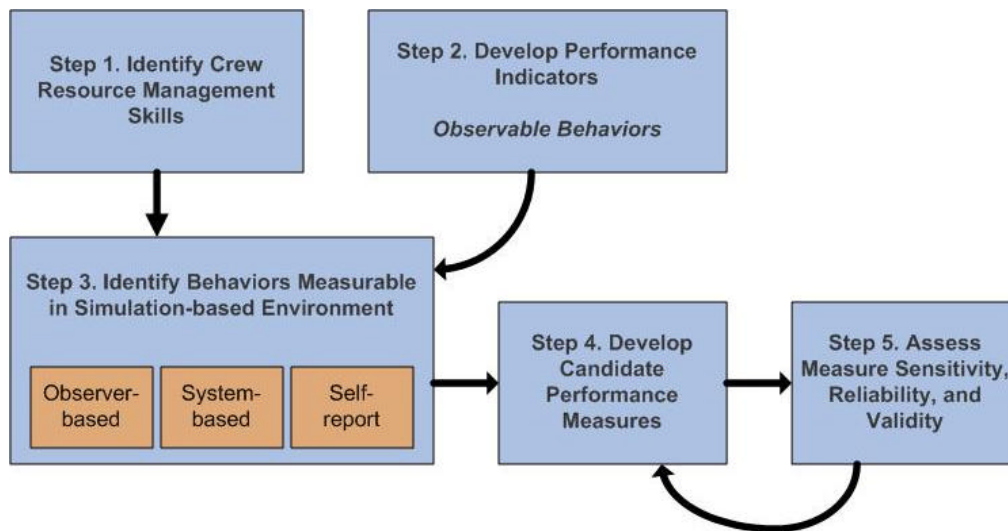


Figure 1. Performance measure development process.

#### Step 1. Define CRM Skills

The first step of the performance measure development process was to review relevant documents in order to define the CRM skills that pilots need to interact effectively with advanced technologies (in this case, SVS and EVS). Several existing definition and measurement structures for CRM behaviors and skills were reviewed, including the University of Texas Line/LOS Checklist (Helmreich, Butler, Taggart, & Wilhelm, 1995), the Approach and Landing Accident Coding Form (Khatwa & Helmreich, 1998), the European NOTECs system (Flin & Martin, 2001), the CRM Assessment System Expert Tool (Dutra, Norman, Malone, McDougall, & Edens, 1995), the Situation Test of Aircrew Response Styles (Hedge, Bruskiwicz, Borman, Hanson, Logan, & Siem, 2000), and a checklist developed by Eduardo Salas and colleagues for CRM training (Salas, Wilson, Burke, Wightman, & Howse, 2006). Because the CRM checklist developed by Salas and colleagues represents one of the most recently published papers on CRM evaluation, this work served as a principal source during definition development.

We modified and updated CRM skill definitions found in the literature to improve their applicability in assessing the influence of technology insertion on performance, and generated new definitions for factors that were not covered by the existing literature. As a result of this process, we generated

seven revised definitions of CRM skills, as shown in Table 1. These CRM skills essentially serve as a framework for observer-based measure development.

Table 1. CRM Skills and Definitions.

CRM Skill	Definition
Communication	Ability of crew members to clearly, concisely, and accurately send and receive information in a timely manner, and to provide useful feedback.
Anticipation & Planning	Ability of crew members to predict likely future states and develop a course of action by organizing resources, activities, and responses to ensure that tasks are completed and synchronized.
Coordination	Ability of crew members to accurately monitor and assess their own and other team members' performance. Ability of team members to sequence, pace, and deconflict activities, and to balance individual workload.
Leadership	Ability of crew members to encourage team members to work together, motivate each other, and establish a positive team atmosphere.
Decision Making	Ability of crew members to gather and integrate information, and to make logical and sound judgments.
Adaptability	Ability of crew members to alter their behavior or strategies based on contingency planning and/or as new information becomes available.
Situation Monitoring	Ability of crew members to develop and maintain an understanding of the task, aircraft systems, and environment.

## Step 2. Identify Performance Indicators

The second step was to conduct a series of knowledge elicitation sessions with subject matter experts (SMEs) and targeted users (NASA researchers) in order to identify likely performance indicators. A performance indicator is an observable behavior that allows an expert rater to recognize whether the crew is performing well or poorly on CRM skills. During this step, it is critical to identify observable behaviors rather than inferred behaviors to develop measures that are less sensitive to individual rater differences and that multiple raters can reliably assess.

We used the Critical Incident Technique (Anderson & Wilson, 1997) to generate multiple scenario options where SVS/EVS technologies might be used. Specifically, we asked the pilot SMEs (four certified flight instructors and professors from Embry-Riddle Aeronautical University) to describe situations in which the insertion of SVS/EVS may mitigate, prevent, or elicit pilot error in terms of CRM behaviors and/or overall performance. The majority of situations described by the pilot SMEs involved taxiing and approach/landing phases of flight. We then constructed three high-level scenarios, based on information provided by the pilots, to use as frameworks for eliciting performance indicators – to initiate discussion regarding specific behaviors pilots might exhibit under varying conditions. These scenarios specified a variety of situations in which SVS/EVS technologies might be used, including airports with terrain-challenging conditions (e.g., ASE: Aspen, CO), approach type (e.g., Category I), time of day (e.g., night), weather conditions (e.g., fog), and unexpected events (e.g., baggage cart on runway).

We asked the pilot SMEs to walk through the high-level scenarios from the perspective of an instructor pilot, and list what CRM behaviors they would observe and how the presence of an SVS/EVS integrated system might influence those behaviors. We used high-level scenarios to elicit CRM behaviors to ensure that the performance indicators were not associated with any specific scenario event (e.g., off-normal events or other scenario-specific details) or technology implementation (e.g., heads-up vs. heads-down displays), but would capture CRM performance in all scenarios. We then met with NASA researchers to finalize the list of performance indicators.

Developing performance indicators was an iterative process that involved several rounds of discussion

among the Aptima team members, pilot SMEs, and NASA researchers. Discussions primarily involved obtaining more detailed information about the observable behaviors as well as identifying the potential sequence of occurrence of these behaviors within a given phase of flight. Table 2 shows sample performance indicators associated with various phases of flight.

Table 2. Example Performance Indicators.

Phase of Flight	Performance Indicator
Takeoff	Configures aircraft appropriately for takeoff
Takeoff	Makes proper callouts
Approach	Maintains appropriate aircraft rate of descent
Approach	Configures SVS/EVS properly
Approach	Makes decision to land/go-around at proper decision height
Missed Approach	Makes appropriate maneuvers to initiate go-around
Taxi to Gate	Coordinates what taxiway to use off runway
Taxi to Gate	Maintains proper navigation on taxiway
Taxi to Gate	Monitors taxiway for traffic

## Step 3. Identify Measurable Behaviors

The third step involved identifying what observable behaviors *should* be measured in assessing the effects of advanced technologies on CRM performance. To understand what should, we applied three decision criteria, namely that the measures must be (1) *measurable in a simulation-based environment*, (2) *CRM-related*, and (3) *sensitive to the usage of SVS/EVS*.

Table 3 presents an excerpt of the decision criteria applied to a set of performance indicators. The “**measurable?**” column asks the question, “Does the performance indicator represent a behavior that can be measured within the simulation-based environment?” Three types of measures are potentially available in a simulation-based environment: (1) data obtained by observation (observer-based measures); (2) data obtained by self-report (self-report measures); and (3) data taken directly from the simulation (system-based measures). If the team (i.e., Aptima human factors scientists, pilot SMEs, and NASA researchers) agreed that the performance indicator described a behavior that could be observed by an expert rater,

we entered “Observer” in this column. If the performance indicator described a behavior that could be rated by the pilot, we entered “Self-Report” in this column. If the performance indicator described a behavior that could be measured by the simulator, we entered “System” in this column.

Table 3. Excerpt from the Performance Indicator and Decision Criteria Matrix.

Performance Indicator	Measurable?	CRM Related?	Likely affected by SVS/EVS?
Configures aircraft and systems appropriately for takeoff	System	No	No
Makes proper callouts (e.g., speeds, configuration changes)	Observer	Yes	No
Maintains appropriate aircraft rate of descent	Observer, System	No	Yes
Configures SVS/EVS properly (one or both pilots, depending on single or dual SVS/EVS installation)	Observer, System	Yes	Yes
Makes decision to land/go-around at proper decision height	Observer	Yes	Yes
Makes appropriate maneuvers to initiate go-around (e.g., turn and climb)	Observer, System	No	Yes
Coordinates what taxiway to use off runway	Observer	Yes	Yes
Maintains proper navigation on taxiway (e.g., follows cleared route on taxi)	Observer	No	Yes
Monitors taxiway for traffic	Observer, System, Self-report	Yes	Yes

Although the current paper only discusses the development of observer-based measures, the combination of complementary data types has the potential to yield a more robust and comprehensive representation of crew performance. For example, system-based data can be used to validate observer-based and self-report data; trained observation can provide insights that are not easily obtained from system-based data; and self-report data can provide information on cognitive factors that are not externally observable.

The “**CRM related?**” column asks the question, “Is the performance indicator capturing behavior that is related to at least one of the CRM skills defined in Step 1?” CRM-related behavior refers back to the seven skills/definitions, shown in Table 1, established for the purpose of this project: communication, anticipation and planning, coordination, leadership, decision making, adaptability, and situation monitoring. If the performance indicator was related to at least one of the seven CRM skills, we placed a “Yes” in this column. If the performance indicator was not related to at least one of the seven CRM skills, we placed a “No” in this column.

The “**likely affected by SVS/EVS?**” column asks the question, “Would the crew's performance on this performance indicator be impacted (positively or negatively) by having SVS/EVS onboard?” If we would expect performance to be influenced by

SVS/EVS usage, we placed a “Yes” in this column. If we would not expect performance to be impacted by SVS/EVS usage, we placed a “No” in this column.

We analyzed these data to identify the “rich” areas for assessment—those observable behaviors that draw most extensively on the relevant CRM skills and are likely to be affected by SVS/EVS.

#### Step 4. Develop Performance Measures

The development of observer-based performance measures entails a considerable time investment as well as knowledge elicitation expertise. We worked with pilot SMEs to develop candidate performance measures with behaviorally-anchored rating scales. Behaviorally-anchored rating scales tie specific, observable behaviors to good, average, and poor performance. Specifically, we developed these measures through a series of structured group interviews with pilot SMEs. We concentrated on those performance indicators that met the decision criteria we established – that is, those performance indicators identified as being (1) *measurable in a simulation-based environment*, (2) *CRM-related*, and (3) *sensitive to the usage of SVS/EVS technologies*.

Prior to the first interview, we developed a number of draft performance measures to serve as starting points for discussion. Our goal for the interviews was to focus on performance measure relevance, observability, and wording along with scale type and anchor wording. We first asked the pilot SMEs if the performance measure wording was specific enough to get at observable behaviors, and changed the wording accordingly in real time. Next, we determined whether a Likert scale was appropriate, or if the measure required only a “Yes/No” response. We then developed the behavioral anchors associated with the Likert scales and asked the pilot SMEs to define good/poor performance in terms of observable behaviors. We used specific questions to help identify these observable behaviors. The questions included: What does the crew *do* or *say* to indicate good/poor performance for this measure? What would *cause* the crew to do well or poorly at this measure? In what *situations* will the crew perform well or poorly on this measure? What behaviors would represent a rating of 1/3/5 along the Likert scale?

The pilot SMEs also helped identify the appropriate sequence for the performance measures. Table 4 shows a subset of the candidate performance measures developed as a function of these interviews.

Table 4. Candidate Performance Measures.

<p>1. Does the crew coordinate to configure the SVS/EVS in a timely manner?</p> <p>1   2   3   4   5</p> <p>Does not coordinate to configure the SVS/EVS   Coordinates to configure the SVS/EVS, but not in a timely manner   Coordinates to configure the SVS/EVS in a timely manner</p>
<p>2. Does the crew run diagnostic tests on the SVS/EVS as the situation requires?</p> <p><input type="checkbox"/> Yes <input type="checkbox"/> No</p>
<p>3. Does the crew recognize and deal with unreliable SVS/EVS information?</p> <p>1   2   3   4   5</p> <p>Does not recognize unreliable information   Recognizes unreliable information, but does not handle appropriately   Recognizes unreliable information and handles appropriately</p>
<p>4. Rate the crew's communication:</p> <p>1   2   3   4   5</p> <p>Non-sterile cockpit, appropriate callouts omitted   Non-sterile cockpit, but all appropriate callouts made   Sterile cockpit, all appropriate callouts made</p>

### Step 5. Performance Measure Testing

The fifth step involves assessing and revising candidate performance measures. As illustrated in Figure 2, this step involves assessing candidate performance measures considering:

1. *Sensitivity*. To be useful in differentiating performance, measures should be sensitive to varying levels of performance. Does the measure distinguish among 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)? Does the measure distinguish among multiple performance levels associated with using SVS/EVS technologies?

2. *Reliability*. In this context we are concerned with inter-rater reliability. For measures that are based on observation, do multiple observers rate the same behavior in the same way? To make comparisons across crews or within a crew over time, when different raters may be assessing performance, any variability in ratings should be due to performance rather than the difference in raters.

3. *Validity*. Because there is no “gold standard” or external performance criterion against which to compare these ratings, a measure of construct validity can be used by comparing the comprehensive mean score on the measures to overall ratings of flight crew performance given by the raters. If these observers agree, then the correlation between this overall rating and the more detailed measures can be used to validate the detailed measures, a “convergence of

experts” validation technique (Holt, Boehm-Davis, & Beaubien, 2001).

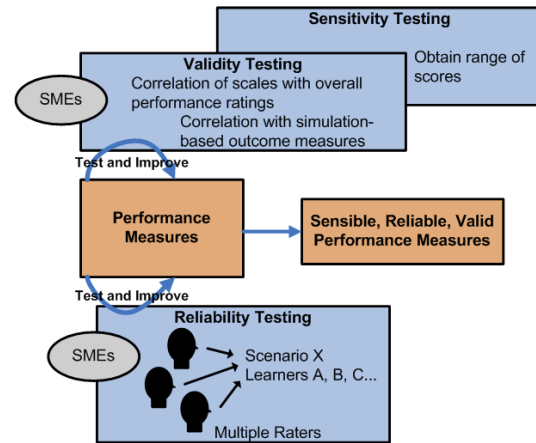


Figure 2. Assessment process for measure sensitivity, reliability, and validity.

This assessment process has not yet been completed for the current set of candidate performance measures. However, the assessment of measures developed for the air-to-air combat domain through a similar process provides a successful use case for measure evaluation related to our current efforts. MacMillan and colleagues (MacMillan et al., in press) had six air-to-air SMEs observe recorded data of ten F-16 four-ship teams, and asked them to independently rate each of the ten teams both in real time and at the end of a scenario based on overall team performance on the mission. The authors found that expert ratings on observer-based performance measures were able to differentiate the teams, showing neither a floor nor ceiling effect. Furthermore, coefficient alphas computed to assess inter-rater reliability showed that the behaviorally-anchored scaled provided a high degree of reliability across observers. Finally, a high correlation between individual measure ratings and overall performance ratings indicated that the performance measures provided valid assessments of overall team performance as well as reliable measures of performance on specific aspects of behavior.

### Conclusions

This research involved the development of observer-based measures sensitive to the insertion of advanced flight deck technologies, such as SVS and EVS. The goal was to provide researchers with the means to make sensitive, reliable, and valid ratings of CRM performance in relation to using SVS/EVS technologies in a simulation-based environment.

The approach we used to develop these observer-based measures can be extended to other advanced technologies, such as electronic flight bags and head-up guidance systems, as well as other domains, including pilot training, air traffic control, and healthcare. For example, meaningful, quantitative measures of crew performance can aid airline training managers in assessing the impact of advanced technology training programs on increased safety-related performance and situation awareness as well as reduced performance time and workload.

As mentioned previously, the combination of complementary data types, such as observer-based, self-report, and system-based measures, has the potential to yield a more robust and comprehensive representation of crew performance. Next steps in this research could include applying this process to the development of self-report and system-based measures, and then developing a methodology to effectively integrate these data sources.

#### Acknowledgments

This material is based upon work supported by the National Aeronautics and Space Administration under Contract No. NNL06AA29P issued through the Aviation Safety Program. Data first produced in the performance of the contract. Lynda Kramer was the NASA Langley Research Center Technical Point of Contact. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NASA. The authors wish to thank Lynda Kramer, Randy Bailey, and Mike Norman for their participation in this research. The authors also wish to thank Ted Beneigh, Tim Plunkett, Pat Donahue, and Inderpreet Singh for serving as pilot SMEs.

#### References

- Alexander, A. L., Wickens, C. D., & Hardy, T. J. (2005). Synthetic vision and the primary flight display: The effects of guidance symbology, display size, and field of view. *Human Factors*, *47*, 693-707.
- Anderson, L., & Wilson, S. (1997). Critical incident technique. In D. L. Whetzel & G. R. Wheaton (Eds.). *Applied measurement methods in industrial psychology*. Palo Alto, CA: Davies-Black Publishing.
- Boeing, (2005). Statistical Summary of Commercial Jet Airplane Accidents, World Operations 1959-2004, May 2005. Retrieved August, 2005, from: <http://www.boeing.com/news/techissues/pdf/statsum.pdf>.
- Cooper, G. E., White, M. D., & Lauber, J. K., (1980). Resource management on the flightdeck. *Proceedings of a NASA/Industry Workshop* (NASA CP-2120). Moffett Field, CA: NASA-Ames Research Center.
- Dutra, L., Norman, D., Malone, T., McDougall, W., & Edens, E. (1995). Crew resource management assessment system: Identification of key behavioral markers. In *Proceedings of the 8<sup>th</sup> International Symposium on Aviation Psychology* (pp. 562-567). Columbus, OH: The Ohio State University Press.
- Flight Safety Foundation (2005). *Priorities*. Retrieved August, 2005, from: <http://www.flightsafety.org/priorities.html>.
- Flin, R. & Martin, L. (2001). Behavioral markers of crew resource management: A review of current practice. *IJAP*, *11*, 95-118.
- Hedge, J., Bruskiwicz, K., Borman, W., Hanson, M., Logan, K., & Siem, F. (2000). Selecting pilots with crew resource management skills. *IJAP*, *10*, 377-392.
- Helmreich, R. L., Butler, R. E., Taggart, W. R., & Wilhelm, J. A. (1995). *The NASA/University of Texas/FAA Line/LOS checklist: A behavioral marker-based checklist for CRM skills assessment* (Technical paper 94-02). Houston, TX: University of Texas Aerospace Crew Research Project.
- Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (1999). The evolution of crew resource management training in commercial aviation. *IJAP*, *9*, 19-32.
- Holt, R. W., Boehm-Davis, D. A., & Beaubien, J. M. (2001). Evaluating resource management training. In E. Salas & C. A. Bowers (Eds.), *Improving teamwork in organizations* (pp. 165-188). Mahwah, NJ: Lawrence Erlbaum Associates.
- Khatwa, R., & Helmreich, R. L. (1998). Analysis of critical factors during approach and landing in accidents and normal flight. *Flight Safety Digest, November 1998-February 1999*, 1-92.
- MacMillan, J., Entin, E. B., Morley, R., & Bennett, W. (Manuscript in press). Measuring team performance in complex and dynamic military environments. *Military Psychology*.
- Prinzel, L. J. III, Comstock, J. R., Glaab, L. J., Kramer, L. J., Arthur, J. J. & Barry, J. S. (2004). The efficacy of head-down and head-up synthetic vision display concepts for retro- and forward-fit of commercial aircraft. *IJAP*, *14*, 53-77.
- Salas, E., Wilson, K. A., Burke, C. S., Wightman, D. C., & Howse, W. R. (2006). The design, delivery and evaluation of CRM training: A checklist. *Ergonomics in Design*, *14*, 6-15.
- Schnell, T., Kwon, Y., Merchant, S., & Etherington, T. (2004). Improved flight technical performance in flight decks equipped with synthetic vision information system displays. *IJAP*, *14*, 79-102.